DOCUMENT RESUME

ED 358 258 CE 063 774

TITLE	Performance Testing Manual and Workbook for Vocational Education Administrators and Teachers.
INSTITUTION	Florida State Dept. of Education, Tallahassee. Div. of Vocational, Adult, and Community Education.
PUB DATE	Apr 91
NOTE	94p.
AVAILABLE FROM	Division of Vocational, Adult, and Community Education, Bureau of Career Development and Educational Improvement, Florida Education Center, Tallahassee, FL 32399-0400 (order no. AD 433 BK 91, workbook; AD 434 BK 91, manual).
PUB TYPE	Guides - Non-Classroom Use (055)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Interrater Reliability; *Performance Tests; Postsecondary Education; Scoring; Secondary Education; Task Analysis; *Teacher Made Tests; *Test Construction; Test Content; Test Format; *Test Interpretation; Test Norms; Test Reliability; Test Use; Test Validity; *Vocational Education; Work Sample Tests

ABSTRACT

This packet contains a manual and a workbook for developing performance tests in vocational education. The manual gives an in-depth description of how to develop, score, and use performance tests. It includes the following sections: definitions of performance testing, steps in developing a performance test, selecting a performance development committee, determining what should be tested, developing a test blueprint, developing the test, training the raters, pilot testing the instrument, determining reliability and validity, setting cut scores, and reducing errors and other issues. An example of how to develop a performance test is provided along with an actual student exercise. Appendixes provide technical information on evaluation methods, including the correlation coefficient, the Phi coefficient, and Cohen's kappa. The workbook is an abbreviated reference source that can serve as a working document for use with committees during performance test development activities. It provides step-by-step directions for implementing the information provided in the manual. Appendixes contain the following: task analysis, a test blueprint form, performance test scoring systems (checklist and rating scale), and a sample performance test item. A 21-item glossary is also included. (Contains 16 selected references.) (KC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED358258

Performance Testing Manual and Workbook for Vocational Education Administrators and Teachers

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

L Meadows

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CE063774

PERFORMANCE TESTING MANUAL:

For Vocational Education Administrators and Teachers

3

April, 1991

Sponsored and published by:

Florida Deparment of Education
Betty Castor, Commissioner of Education

For further information contact:

Instructional Support Services Section
Bureau of Program and Staff Development
Division of Vocational, Adult, and Community Education
Department of Education
Florida Education Center
Tallahassee, FL 32399-0400
(904) 487-3750/Suncom 277-3750

DVACE/04/91/

TABLE OF CONTENTS

Introduction 1
Definition of performance testing 1
Steps in development of a performance test 2
Select a performance development committee 2
Determine what should be tested 2
Identify critical skills to be tested 2
Consider time available for testing 2
Consider resources available for testing 3
Break down complex skills 3
Avoid measuring attitudes 3
Consider test difficulty 3
Example 4
Summary 4
Task Analysis Worksheet 5
Develop test blueprint 6
Definition 6
Test length 6
General scoring procedures 7
Weighting of different skills 8
When to test 8
Conditions of testing 9
Who may serve as a rater 9
Test security 9
Summary 10
Blueprint form 11
Develop the test 13
Design test scenario 13
Write Instructions 13
List equipment and supplies 13
Develop rating scales and/or checklists 13
Rating Scales 13
Checklists 16
Summary 18
Train the raters 18
Necessity of training raters 18
Common rater errors 18
Summary 19
Pilot test the instrument 19
Determine reliability 19
Definition 19
Test-retest reliability 19
Inter-rater reliability 19
Intra-examiner reliability 20
Factors influencing reliability 21
Summary 21
Determine validity 21
Defintion 21
Content validity 22
Criterion-related validity 22
Construct validity 22

Validity and performance tests 23
Establishing content validity 23
Establishing criterion-related validity 23
Establishing construct validity 23
Summary 24
Set Cut Scores 24
Introduction 24
Categories of cut score methodologies 24
Angoff method 25
Example with equally weighted items 25
Example with differentially weighted items 26
Modifications of the Angoff method 27
The Ebel method 27
Example 28
Methods based on judgments about individuals 29
Example 29
Other methods 30
Summary 30
Other Issues 31
Summary 31
Example 32
Actual Item 45
References 48
Other readings of interest 49
Appendix 50
Correlation coefficient 51
Phi coefficient 52
Cohen's kappa 54
Glossary 56

INTRODUCTION

As many definitions of performance tests exist as there are authors writing about them. Some definitions are quite technical, while others are relatively straightforward. However, a common theme runs through them all. For the purposes of this manual, **a performance test is defined as any assessment procedure which simulates the behavior a student would display on the job with as much realism to the actual job behavior as possible.**

There are three primary types of performance tests. The first, direct assessment, involves the direct observation of a performance on the job. Direct assessment is not practical for the classroom setting.

The second type of performance test is a work sample. This occurs when a specified task is observed in a controlled setting (such as a classroom). The tools, materials, and other resources available to the student must be similar to those on the job and the tasks the students perform must be equivalent to real job tasks.

The third type of performance test is a simulation. Simulations involve abstracting some aspect of the real job. Examples would include providing pilots with a simulation of a cockpit in which to learn to fly rather than allowing them to fly a real plane; providing medical students with cadavers on which to practice surgical skills; auditory or visual stimuli (for example, asking health students to interpret recorded heart sounds), and role playing. In general, simulations will be useful whenever use of the real equipment could be dangerous, when errors could have serious consequences, when the actual stimulus may be infrequent or difficult to obtain, or whenever some type of oral response is required (such as with role playing).

There may be some degree of overlap between work samples and simulations, so the "type" of performance test one is employing for a given purpose may not be easily categorized. In general, the more closely the assessment matches the actual job performance, the better the assessment procedure will be. Thus, when feasible, work samples would be somewhat preferable to simulations. However, simulations can be quite realistic (such as substituting inert chemicals for possibly toxic ones when asking a student to mix or apply such chemicals) and, thus, are just as useful as work samples.

There are a number of steps to follow when developing a performance test. Every step in the process of test development should be clearly and comprehensively documented in case of legal action. The following steps can be used to develop a performance test:

1. Select a performance test development committee.
2. Determine what should be tested.
3. Develop the test blueprint.
4. Develop the test.
5. Train the raters.
6. Pilot test the instrument.
7. Determine reliability.
8. Determine validity.
9. Set cut scores.

Each of these steps will be discussed in further detail.

SELECT A PERFORMANCE DEVELOPMENT COMMITTEE

The first step in developing a performance test is to select a committee to determine what is to be tested and to develop the test. Committee members should include teachers, local industry representatives, or program advisory committee members, and a testing and measurement specialist, if available. Committee members should be subject matter experts in the area in which they are to develop the test. Thus, auto mechanics teachers and local automechanic technicians should serve on the committee to develop the performance test for automotive mechanics.

The issue of who determines the material to be included on the test may have legal implications, so it is important that content experts determine the skills to be covered on any test, regardless of its type.

DETERMINE WHAT SHOULD BE TESTED

Identify Critical Skills to be Tested

Tasks selected for the test should evaluate different critical competencies and should assess specific skills required on the job. A representative sample of the most critical tasks should be included. It is usually best to include several shorter tasks rather than one or two longer ones. This may increase reliability and will increase the representativeness of the test. In choosing tasks to be tested, consider the consequences of allowing a student to remain deficient in a particular skill. If the student can perform the job without the skill, another skill which is of critical importance to job performance should be substituted. The tasks to be included on the test may be drawn from those in the curriculum frameworks. Several student performance standards may be tested with a single performance example (see the example given at the end of this manual).

Consider Time Available for Testing

One very practical consideration is the time available for testing. In general, performance tests can last anywhere from two to sixteen hours or even longer. Disregarding the cost of

materials consumed during the performance test (for example, paint, chemicals, or flour), the longer the performance assessment, the more expensive it is, due to the time required of raters. On the other hand, the longer the time available for testing, the greater the number of tasks that can be measured resulting in a more representative assessment. Generally, two hours would be a minimum testing time for the purposes of the Gold Seal Endorsement.

Consider Resources Available for Testing

Resources such as personnel, time, and the equipment needed should be considered. Alternative methodologies of assessment which are less costly should be used if possible. For example, identification of tools can be easily tested using paper and pencil methodologies. Paper and pencil methodologies are less expensive because they allow for group administration and do not require two or more raters per student.

Break Down Complex Skills into Measureable Components

Complex skills, such as "Check two-wheel alignment", should be broken down into their basic, identifiable and measureable components through the process of instructional task analysis. A sample task analysis worksheet is provided at the end of this section. These components should be defined in terms of what the student actually has to do. Begin by asking basic performance questions about the task to be assessed. For example, "What do you have to do to check two-wheel alignment?" These components will become the basis for rating scales or checklists that will be used and will play an important role in the development of rating scales and/or checklists for scoring.

Avoid Measuring Attitudes

It is best not to test personal values or attitudes unless it can be empirically demonstrated that these characteristics are related to successful job performance. Interpersonal skills, such as those required for employability skills, may be tested using a role play. However, role playing is usually best for training, rather than assessment. It is also expensive because it requires at least three observers per student--two raters and one interviewer.

Consider Test Difficulty

Another important issue in deciding what to test is the difficulty of the test. Assess only the skill level which must be obtained for entry-level job performance.

Example

At the end of this manual, an example of the development of a performance test will be provided to help clarify some of these concepts. This example will focus on the task, "Check two-wheel alignment on an automobile".

The illustrative task of "checking the two-wheel alignment on an automobile" would be only one of the tasks to be included on a performance test. Other tasks selected for inclusion might be:

- Reface valves and seats.
- Inspect front suspension and steering systems.
- Perform standard tests on engine electrical system.
- Inspect and evaluate front disc brakes.
- Pressure test the air conditioning system for leaks.

In addition, a second volume accompanies this manual. This second manual is a workbook which provides a brief review of the material covered in this manual.

Summary

When choosing tasks for inclusion on a performance test, the following steps must be taken:

- Remember to involve content experts in selecting tasks to be included on the performance test.
- Identify the most critical skills for possible inclusion on the performance test.
- Consider the time available for testing and the time required for the tasks under consideration.
- Consider resources available and other assessment methodologies.
- Select a representative sample of critical tasks for inclusion on the performance test.
- Break down complex tasks into their basic components.
- Select tasks appropriate for entry level job performance.
- Avoid testing personal values or attitudes, if possible.
- Consider test difficulty.

TASK ANALYSIS

Instructional Task Analysis is a procedure that requires both logic and creativity. It takes your knowledge as subject-matter experts to extract the steps of performing a particular job task. It takes creativity to accurately describe those steps.

After you have completed the anlysis of a task, the simple "who, what, when, and how" check can help assure that you have covered most of the information needed:

1. Is **who** does the task relevant and clearly stated or implied?
2. Is **what** is needed to perform the task successfully identified and accurately described?
3. Is **when** the task is applied and completed relevant and clearly stated or implied?
4. Is **how** the task is performed described accurately and in appropriate terms?

Procedures for Analyzing Tasks

1. Identify the **Performance skill** from the representative skills selected for the performance test.

2. Determine the **Standard of Performance**, i.e., what level of performance denotes that the task has been satisfactorily completed.

3. Identify the **Conditions for Performance** of the task, e.g., tools, equipment, supplies, environment.

4. Identify step-by-step what a person must know and do in order to perform the task.

5. Develop the **performance rating scale** or **performance cheklist** for the task based on the analysis of the performance steps of the task.

DEVELOP TEST BLUEPRINT

Definition

A blueprint is a plan for a test. The test blueprint will help ensure that the test meets its purpose. Once the blueprint is developed, the test itself will be quite easy to develop. The test blueprint usually contains information about test length, scoring methodologies, weighting of different skills, when to test, the conditions under which testing is to occur, who can serve as a rater, and test security. Each of these will be discussed further. A test blueprint form is provided at the end of this section and a sample blueprint is provided in the illustrative example at the end of this manual.

Test Length

Test length is a function of both practical and measurement considerations. Tasks included on the test usually determine how long the test takes. Considering the time available for testing is instrumental in determining the tasks chosen for the test and will be of importance in determining test length.

The greater the number of tasks that can be included on the assessment, the more representative it will be. Greater representativeness may result in greater reliability and validity. Thus, the more tasks that can be included the better. However, the test should only be as long as needed to reliably measure the skills that are on the test. Thus, test length for different occupational program areas will vary.

In order to include more tasks, you can determine if a time limit can be imposed on the tasks (when such a time limit is not normally an aspect of the task). For example, time would not normally be part of the scoring for checking tire alignment of an automobile. However, if most students can finish this task in one hour, it would seem undesirable for a student to take a great deal longer, say two hours. If no time limit is imposed, students should not be penalized for working slowly unless time and speed are essential for successful performance (for example, a typing test).

Another consideration in determining test length will be what types of decisions will be made from the test. The performance tests developed for the Gold Seal Endorsement will be used to make relatively important decisions about students. Thus, more evidence will be required to obtain acceptable levels of reliability and validity. This may require a longer test.

General Scoring Method

There are several issues which must be decided when considering how a performance test is to be scored:

1) Must the student pass every skill or just a certain percentage?
2) Are there certain tasks within a skill that the student must pass?
3) Will the student be scored by skill or will one overall score be assigned? This is related to analytic vs. holistic scoring, discussed below. If the student is to be scored by skill, these individual skill scores must be reliable and valid.
4) How will each skill be scored (e.g., a rating form or checklist)? These will be discussed more fully in a later section and examples of both a rating form and a checklist are provided.
5) Will the process or product be scored? This issue is discussed in greater detail below.

It must be determined if holistic or analytic scoring will be used. In analytic scoring, each criterion or component of a task is evaluated individually and assigned a score. Then, the component scores are summed to produce an overall score. In holistic scoring, all criteria are considered simultaneously and only an overall score is assigned. Even with holistic scoring, some criteria can be given more weight than others. Holistic scoring is less time consuming and less expensive but doesn't provide as much information. Analytic scoring can provide diagnostic information for remediation if that is appropriate.

Another issue which must be considered is whether or not the process, the product, or both will be scored. The product is a tangible outcome of performance (such as a cake), while process refers to the methods used to arrive at a particular point. If the process is to be scored, it must be assessed at the time of performance while a product may be analyzed at a later time. The process may or may not yield a product. Process is often more difficult to assess due to its subjective and/or complex nature. If the essential aspect of the skill is the process, analyze it. If the product is more important than the process, the product should be analyzed. However, both may be scored. Process is usually assessed with a rating scale or checklist, while the product may be assessed with a rating scale, checklist, or a product scale (a scale which presents representative samples of products arranged in order of quality and the student's score is assigned on the basis of the product criterion it most closely resembles). The choice of process or product evaluation depends on the skill to be measured and practical considerations. For example if there is no tangible product, the process must be scored. Some researchers (e.g., Slater, 1980) suggest that it is more important in student certification to use product evaluation because the goal of the assessment is to determine if the student has achieved the learning objectives. Product

evaluation is also less costly, yields more reliable scores, and can be judged by multiple raters and described in objective terms. Slater also notes that it is often more difficult to specify the appropriate steps leading to a given product than to specify the desirable characteristics of the product. Thorndike (1982) also indicates that product evaluations have certain advantages. For example, product scoring yields a permanent record which can be examined slowly and in as great a depth as desired. A uniform scoring procedure can be applied yielding a relatively high standard of reliability. In addition, several components of the product can be evaluated (e.g., the time it took, the level of perfection, etc.). Thorndike notes, however, that a detailed scoring guide is needed in order to be sure that the same standards are applied to all student products. The guide should specify the elements to be scored and scoring standards for each element.

Other researchers, such as Livingston (1980) and Klein (1980) suggest that both process and product be evaluated, particularly when a bad process does not always result in a bad product. Klein (1980) indicates that the National Occupational Competency Testing Institute (NOCTI) uses at least four product and four process tasks to obtain an acceptable reliability.

Once these issues have been decided, the actual scoring procedure can be specified. Regardless of the decisions regarding analytic or holistic scoring or process or product scoring, clear and unambiguous standards for scoring must be established. Performance tests can be scored by counting the number of correct items if there are enough items on the test. Certain skills can be weighted more heavily than others if that is appropriate.

Weighting of Different Skills

Weights may be assigned to the tasks to reflect the importance and frequency with which the tasks are performed on the job. The more important the tasks, the higher they should be weighted. The more frequently they occur on the job, the higher they should be weighted. The example at the end of the manual provides an example if differential weighting. Differential weighting of tasks is optional.

When to Test

The issue of concern here is whether to use an on-going assessment or an end of the program assessment. On-going testing usually involves observing everyday events as they occur, while an end of the program assessment usually involves designing specific exercises which will elicit the desired performance. Of course, both may be used in combination. If the behavior occurs infrequently or you are making very important decisions about the students (as for the Gold Seal Endorsement), it is wise to use structured exercises. Test security is usually of greater concern when using an on-going assessment.

An issue related to whether to use an on-going or end or program assessment is the obtrusiveness of the measurement. If an on-going assessment is chosen, the measurement is usually unobtrusive, as the student may not even be aware that he or she is being evaluated. The end of the program assessment is more obtrusive. If an on-going, unobtrusive assessment, is chosen be sure that periodic observations that yield a fair and representative sample of performance are made. If you use an end of program assessment, be sure that the situation is standardized and observations are planned carefully.

Conditions Under Which Testing is to Occur

The directions given to the student, conditions under which testing occurs, and equipment and materials needed for each performance item must be specified carefully. This allows for standardization of testing. Tests given under different conditions using different equipment or using different materials, will not yield comparable results. All students must be given the same task under the same conditions and be scored in exactly the same way. All equipment should be in the same operating order and calibrated in the same manner for all students.

If the student is unfamiliar with a particular piece of equipment or some of the materials used in the test, he or she must be given the opportunity to become familiar with the equipment or material. Students should be given the opportunity to run unfamiliar equipment for a short while before testing occurs.

Who May Serve as a Rater

Generally, more than one rater will be needed to observe and rate student performance on the performance test items. Who may serve as a rater will have important implications for reliability of the test. The qualifications required of test administrators and raters must be clearly specified. Test administrators and raters must be experienced in the particular field which is being tested. Teachers and industry representatives usually serve as test administrators and raters. In general, however, the student's own teacher should not serve as a rater for that particular student.

Test Security

Test security is not usually an important issue for performance testing since each student must perform the task with a certain degree of proficiency regardless of whether or not he or she knows what will be tested. In fact, most researchers agree that the criteria and procedures for testing should be explained to each student in advance of the test. Students should be informed of the requirements for the Gold Seal Endorsement when they enter the program. They should also be told the skills and objectives to be measured by the test. These skills must then be taught. Both raters and students should be informed of the dimensions to

be evaluated and the purposes of the test before testing occurs.

Summary

When developing a test blueprint, the following elements must be addressed:

- test length
- scoring methodologies and criteria
- weighting of tasks to be scored
- when testing will be conducted
- standardization of the conditions under which testing will occur must be specified
- who may serve as a rater
- a plan for test security

TEST BLUEPRINT FORM

1. Determine Test Length

   _____ Tasks

   _____ Hours of testing time

2. Scoring

   a. Must student pass <u>every</u> skill? (Yes/No) _____

   b. Are there certain tasks the student <u>must</u> pass? (Yes/No) _____
      Skills:

      _____
      _____
      _____
      _____
      _____

   c. Analytic vs. holistic scoring? _____

3. Differential Weighting of skills? (Yes/No)

	Skill	Points
--	_____	____
--	_____	____
--	_____	____
--	_____	____
--	_____	____
--	_____	____
--	_____	____

4. When to test? _____

5. Conditions under which testing occurs.

   _____
   _____
   _____
   _____
   _____

6. Who may serve as raters? (Teachers/Industry Representatives)
   Rater 1: _____
   Rater 2: _____
   Rater 3: _____

7. Test security required? (Yes/No) _____
   If yes, list required security precautions:
   _____
   _____
   _____

DEVELOP THE TEST

Design Test Scenario

Once the test blueprint has been developed, the test itself will be easily generated. In developing the test, create a realistic scenario which requires the student to perform the skills selected for the test.

Write Instructions

Instructions to the students, the raters, and the test administrators should be made explicit. The instructions should clearly state what is required of each person and the criteria for performance. If accuracy is important, this should be made explicit to all parties involved.

List Equipment and Supplies

All equipment and materials required for the performance of each item included on the test must be listed. This list should be as comprehensive as possible.

Develop Rating Scales and/or Checklists for Scoring

The rating scale or checklists to be used for scoring should be developed at this point. Rating scales usually ask raters to place the individual's performance along a continuum, while checklists usually ask the raters to indicate "yes" if the student exhibited that behavior and "no" if the student did not demonstrate the behavior. When you wish to rank order the performance of individuals, ratings of performance, rating scales, or a highly detailed checklist will allow for ranking. When absolute mastery decisions are necessary, a checklist is sufficient. Checklists require less judgment on the part of raters and reduce errors. They also provide information for diagnosing student's weaknesses and documentation should scores be disputed. Regardless of which is used, they should include all dimensions to be rated. If a rating scale is to be used, all the points along the continuum should be described in specific observable terms. Labels such as "poor" to "excellent" will not suffice. An example of a checklist is shown on page 14 and a sample rating scaale is shown on page 15.

Rating Scales

Notice on the sample rating scale, that each number is clearly and unambiguously defined. For example, if the student were to correct the camber (Task 3) to within 2 degrees of the manufacturer's specification, he or she would receive a 2. If the student centered the steering wheel to within 1 degree from center (Task 5), he or she would receive a 4. Numbers are redefined only when the criteria for performance change.

PERFORMANCE CHECKLIST

Directions: Observe the student as he or she performs the task. As the student performs each task, place a check (✓) or X in the correct column (Yes or No).

		Yes	No
1.	Measure and correct the ride height to within 1/2 inch of manufacture's specification.	_____	_____
2.	Read and correct the caster to within 1 degree of manufacturer's specification.	_____	_____
3.	Read and correct the camber to within 1/2 degree of manufacturer's specification.	_____	_____
4.	Read and correct the toe in to within 1/4 degree of manufacturer's specification.	_____	_____
5.	Center the steering wheel to within 2 degrees of manufacturer's specification.	_____	_____

PERFORMANCE RATING SCALE

Directions: Observe the student as he or she performs the task. As the student performs each task, circle the number on the rating scale which indicates the accuracy with which the task is completed.

Task 1: Read and correct the ride height.

Error of more than 1.5" from manufacturer's spec.	Error of 1.1" to 1.5" from manufacturer's spec.	Error of .5" to 1" from manufacturer's spec.	Error of less than .5" from manufacturer's spec.
1	2	3	4

Task 2: Read and correct the caster.

Error of 4 degrees or more from manufacturer's spec.	Error of 3 degrees from manufacturer's spec.	Error of 2 degrees from manufacturer's spec.	Error of 1 degree or less from manufacturer's spec.
1	2	3	4

Task 3: Read and correct the camber.

Error of 3 degrees or more from manufacturer's spec.	Error of 2 degrees from manufacturer's spec.	Error of 1 degrees from manufacturer's spec.	Error of 1/2 degree or less from manufacturer's spec.
1	2	3	4

Task 4: Read and correct the toe-in.

Error of 2 degrees or more from manufacturer's spec.	Error of 1 degrees from manufacturer's spec.	Error of 1/2 degree from manufacturer's spec.	Error of 1/4 degree or less from manufacturer's spec.
1	2	3	4

Task 5: Center the steering wheel.

Error of more than 10 degrees from center	Error of 6 to 10 degrees from center	Error of 2 to 5 degrees from center	Error of less than 2 degrees from center
1	2	3	4

There are two commonly used types of rating scales. The first, like the one in the sample on page 16, is based on numerical ratings (e.g., from 1 to 4), while the second is based on graphic ratings. Numerical rating scales are expressed in the form of numbers. Each number should have constant meaning for each task. For example, if "one" is used to indicate unacceptable performance for one task, it should also indicate unacceptable performance for all other tasks. An example of a very basic numerical rating scale would be:

Center the steering wheel.	1	2	3	4

In this example, there are only numerical points along the scale. However, these numbers are not defined, so it is difficult to unambiguously determine their meaning. Whenever possible, verbal anchors should accompany a numerical rating scale as in the example on page 15. Verbal anchors provide a brief description of each level of performance. Poor examples of rating scales would include both the one above and the one below.

Center the steering wheel.	poor 1	moderate 2	good 3	excellent 4

This latter rating scale is a little better than the first, but the numbers are still not defined in an unambiguous manner, and all raters may not interpret the verbal anchors in exactly the same way. For example, let's say a student centers the steering wheel to within 5 degrees of center. One rater may assign that a value of "good", while another may assign a value of "moderate". The more specific the verbal descriptions, the more likely raters are to agree on what score should be assigned to the student. Look again at the example using specific, behavioral anchors on page 15. This scale is a little longer, but there is no question as to what score a student should receive. When a scale such as this one is used, inter-rater reliability should be quite high.

Graphic rating scales use words as ratings rather than numbers. Graphic rating scales may be quite subjective and may have low reliability if the verbal descriptions lack specificity. For example, the difference between "neat" and "fairly neat" is subjective and will vary from one person to another. Verbal descriptions such as poor, moderate, good, and excellent lead to subjectivity and variation among raters. Use of specific, well defined criteria make the scales less subjective and increase inter-rater reliability.

Checklists

There are two commonly used types of checklists. The quantitative checklist assesses whether or not the task was performed or not. The qualitative checklist attempts to assess the quality of the performance using a yes/no format. In addition, a checklist may be used to record frequency of behavior. A sample checklist is shown on page 14 and a second is

provided in the example at the end of this manual.

There is little evidence to suggest the superiority of either a checklist or a rating scale for use in performance testing. Which is used depends upon personal preference, practical considerations, the type of decision to be made, and the particular behavior to be assessed. Regardless of which is used, however, scoring procedures must be described in detail in order to minimize the chance of a scoring error. In addition, there should be a space for evaluator comments to serve as an internal control. Comments should be consistent with the scoring. For example, if the comments indicate that the student failed to perform certain critical elements which are necessary for passing a skill, yet the skill was checked or rated as passing, a scoring error is likely.

There are a number of steps to follow in order to develop a checklist or rating scale. First, select the number of scoring options for use if a rating scale or product scale is used. Using more than nine dimensions usually results in information of questionable utility. Raters are usually not able to make such fine discriminations. Four to seven options are usually considered adequate. You must also decide if a central point or "uncertain" area is to be part of the scale. If not, the scale should have an even number of scalepoints. When deciding on the number of options, be sure that the options cover the entire range of possible observations, and that the options are mutually exclusive. There should be no overlap (for example, "good to excellent" and "fair to good" is an example where there is overlap). If the options are behaviorally defined, there will usually be no overlap.

The second step in the development of the rating scale or checklist is to list important elements for each task to be performed. This should have been done when developing the blueprint. Each of these skills is a dimension on the rating scale or checklist. Be sure to sequence these tasks in the order in which they will most likely be performed. For example, when "Checking two wheel alignment", the following steps are taken:

- Measure and correct the ride height.
- Read and correct the caster.
- Read and correct the camber.
- Read and correct the toe-in.
- Center the steering wheel.

It is also helpful to include common errors or undesirable options.

Summary

The following steps are to be followed in developing the test:

- Design the test scenario
- Develop the instructions to the test administrators, raters, and students.
- List all equipment and materials required for the performance of each item included on the test.
- Develop rating scales and/or checklists used for scoring
    a. Select the number of options for a rating scale.
    b. Include all important tasks to be performed.

TRAIN THE RATERS

Necessity of Training Raters

All raters must be trained to rate student performance. Priestley (1982) states that

> It is essential that a scorer or observer be trained carefully and thoroughly, both in the specific field or subject area and in how to conduct specific observational assessment. p. 127

Common Rater Errors

There are three common rater errors made when rating student performance. These are personal bias, the halo effect, and logical errors. Personal bias is the observer's tendency to rate all students at the same point along the rating scale. This includes severity errors when judges tend to be too strict in ratings and leniency errors when judges rate everyone high. It also includes judges who rate everyone in the middle of the scale. The halo effect occurs when the rater's general impression of a student causes the rater to rate the person the same on all dimensions. A logical error occurs when the rater mistakenly believes more or less of a relationship exists between two dimensions than actually exists. For example, the rater may think that two things are related and that if a student scores high on one, they will automatically score high on the other. Specifically, let's imagine that "check front disc brakes" and "check rear disc brakes" are included on a performance test. The rater may assume that if the student can check the front disc brakes, the student can also check the rear disc brakes. However, this may not be a valid assumption. In addition to these errors, the possible problem of sex or racial bias may cause raters to rate students in an unfair manner.

The careful training of raters, development of observational rating scales or checklists, and an analysis of inter-rater reliability, discussed below under reliability, will usually help solve most scoring problems. The training of raters should include demonstration and practice. The raters should take the

test while in training to be familiar with some of the problems students may encounter.

Summary

- Train raters in how to use rating scales and checklists.
- Train raters in the scoring methodologies used for the performance test.
- Inform raters of potential rater biases and train them to avoid these.
- Make specific behaviors that the rater must observe explicit.

FIELD TEST THE INSTRUMENT

Field testing the instrument is cricial for insuring that the test has been constructed properly, that all instructions to students, raters, and test administrators are clear, that scoring criteria are clear, time limits set are reasonable, etc.

While the conditions under which the test is actually administered are highly standardized, the instrument may be field tested under a variety of conditions. This will help to identify any areas which need to be clarified.

DETERMINE RELIABILITY

Definition

Any test must be valid and reliable. Legal requirements may require test developers to prove that a test is job-related, valid, reliable, and free from bias. The key question for reliability and validity is how critical and irreversible is the use of the test data. Decisions made in the Gold Seal affect students in important and perhaps permanent ways. Any test results must be valid and reliable. Reliability is the degree to which a test measures a characteristic consistently.

Test-Retest Reliability

Reliability may be established in several ways. However, for performance tests, only two methods of establishing reliability are relevant. The same instrument may be administered after a period of time has elapsed and the student's scores on the two testings correlated. This is called test-retest reliability. The period of time elapsed must be long enough to allow for the effects of memory to diminish. An example of how to compute a test-retest correlation coefficient is provided in the Appendix.

Inter-Rater Reliability

The second method of establishing reliability of a performance

test is inter-rater reliability or how well different raters agree on the score a single performance should receive. Variability among raters arises from two major sources. First, variability within the individual, and second, systematic rater biases (e.g., the halo effect). Simon (1978) suggests six things that will help alleviate variability among raters:

1. training of raters,
2. detailed specifications of the tasks the rater is to perform,
3. providing specific written instructions for raters,
4. reporting information as soon after the observation as possible,
5. using mechanical devises (e.g., videotaping a performance) whenever possible, and
6. getting information from several raters who observe at the same time.

Livingston (1980) indicates that inter-rater reliability is the most important type of reliability for a performance test. Klein (1980) also indicates that an index of inter-rater reliability should be obtained. Livingston suggests at least two raters observe the student at the same time unless the performance is recorded (e.g., videotaped).

There is no consensus on a formula to determine inter-rater reliability. There are three common methods used to estimate inter-rater reliability. The first involves correlating the scores assigned by the two raters. An example of how to compute a correlation coefficient is provided in the Appendix. The second method involves dividing the number of agreements by the number of agreements plus the number of disagreements (total number of ratings). An example is provided in the Appendix. The third method of computing inter-rater reliability is Cohen's kappa (Cohen, 1965). Kappa provides an estimate of agreement between two raters which is corrected for chance agreement. An example is provided in the Appendix. The method you use for calculating interrater agreement will depend to some extent upon the scoring methods used and personal preference.

Intraexaminer Reliability

Another aspect of reliability that is important for performance tests is intraexaminer reliability, or how consistently the rater rates a particular performance across time. Rater's ratings should remain constant over time. Factors such as fatigue and comparison effects may lower the rater's intraexaminer reliability coefficient. This type of reliability should be assessed during rater training. It may be established by asking raters to rate taped performances on several occasions. The performances should represent the range of possible performance, from failing to passing. Special attention may need to be paid to borderline cases, as these may be more difficult to rate than more obvious failures and passing performances.

Factors Influencing Reliability

Reliability on a performance test is greatly affected by factors limiting a rater's ability to rate accurately, such as lack of opportunity to observe, ambiguity of the characteristic being observed (e.g., "leadership"), lack of a uniform standard of reference on the rating scale or checklist, and rater biases such as the halo effect. These problems can, of course, be overcome to a great degree by planning the test administration, defining clearly what is to be observed, specific definitions of rating criteria, and training of raters.

Summary

Reliability is the extent to which a test measures a characteristic accurately and consistently. The following issues are important in assessing reliability:

- Determine the types of reliability to be established. Usually inter-rater reliability and test-retest reliability are used for performance tests.
- Determine test-retest reliability by administering the same items at different points in time and correlating the scores. Be sure to allow enough elapsed time to minimize the effects of memory.
- Determine inter-rater reliability by:
    a. correlating the scores assigned by two or more raters.
    b. dividing the number of agreements by the total number of ratings.
    c. Cohen's kappa.
- Maximize reliability by:
    a. training raters.
    b. giving detailed specifications of the task the rater is to perform.
    c. providing specific written instructions for raters.
    d. reporting information (rating the student) as soon as possible after the observation.
    e. using mechanical devises if possible.
    f. getting information from several raters who observe at the same time.

DETERMINE VALIDITY

Definition

Validity is the most important feature of a test. Messick (1989) defines validity as

> ...an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences

and actions based on test scores or other modes of assessment. p. 13

It is not the test itself which is to be validated, but rather the inferences derived from the test. Validity is a matter of degree, not an all or nothing concept.

Historically, validity was divided into three or four different "types." However, as the definition provided above indicates, validity is now viewed as a single, unitary concept. The traditional "types" of validity are now called categories and the Standards for Educational and Psychological Testing (1985) indicate that the best validation includes several types of evidence from all three of the traditional categories.

Content Validity

Despite the current emphasis on validity as a unitary concept, distinction can still be made between the categories of validity. Content validity is based on professional judgments about the relevance of the content of the test to the content of the domain of interest (e.g., entry level job behaviors) and about how representative the items on the test are to the entire domain. In other words, content validity involves an evaluation by content experts that the test adequately and fairly represents the content covered in the instruction and/or tasks performed on the job.

Criterion-Related Validity

Criterion related validity is based on the degree of empirical relationship between the test scores and criterion scores. This relationship is usually indicated by correlation coefficients or regressions. A distinction is sometimes made between two types of criterion related validity. The first is predictive validity. This indicates how well the test predicts future behavior. The second is concurrent validity. This indicates how well the test predicts current behavior. This category of validity presupposes that the criterion is relevant and has been accurately measured.

Construct Validity

Construct validity is defined as the degree to which the test measures the underlying trait it is supposed to measure. Historically, construct validity was established by looking at the patterns of relationships among item scores or between test scores and other measures. Recently, however, studies have focused on performance differences over time, across groups and settings, and in response to experimental treatments. Both content and criterion related validity are subsumed by construct validity.

Validity and Performance Tests

Livingston (1980) indicates that content validity is the most relevant for performance tests. However, if the purpose of the assessment is to make inferences about student performance in situations unlike those on the test, criterion related validity may be more relevant.

Establishing Content Validity

In order to demonstrate content validity, it must be shown that the behaviors demonstrated on the test are a representative sample of behaviors on the job or that the test produces a product that is representative of the work product of the job. If it is not feasible to demonstrate content validity, the more important the need to demonstrate other types of validity. Including content experts in determining what should be tested will help to establish content validity from the start of the testing process. Once the tasks have been chosen, additional content experts may be asked to review the selected tasks to be certain that they compose a representative sample of critical tasks which an entry level worker would be expected to perform on the job.

Establishing Criterion-Related Validity

In order to establish criterion related validity, one could correlate the student's scores on the performance test with a number of criteria. For example, supervisory ratings, if available, are often used as criteria for establishing criterion related validity. Other criteria may include teacher's ratings, Grade Point Average (GPA), correlations with other tests which measure similar tasks (ideally another performance test), and the like.

Some authors such as Messick (1989), indicate that content and criterion related validity are no longer enough. They contend that construct validity must also be demonstrated.

Establishing Construct Validity

There are two primary ways to establish construct validity. The first is to determine if the test differentiates between those who rank high on the trait the test measures and those who rank low on the trait. For example, if a student who is known to be proficient at a task and one who is known to be much less proficient at the task are given the test, the two should receive scores which would suggest their relative ranking. The student known to be proficient should receive a relatively high score, while the student who is not proficient should receive a relatively low score. The second method of establishing construct validity is to demonstrate that if certain modifications of the trait suggested by theory are made then these modifications are reflected in corresponding test score changes. Demonstration of construct validity in this manner is

somewhat more difficult to apply in many educational settings. For example, consider a new drug rehabilitation treatment, say primal scream therapy. Let's suppose that those who engage in primal scream therapy should kick their drug habits faster than those who engage in other forms of therapy. We can test this theory and, perhaps, establish construct validity by manipulating the type of therapy provided to drug abusers, and observing how fast they are able to kick their habit.

Summary

In summary, the following steps may be taken to establish the validity of a performance test:

- Determine the type of validity to be established.
- Content validity may be established by the selection of appropriate items for inclusion on the test.
- Criterion related validity may be established by administering the test to current students and correlating their scores with external criteria such as supervisory ratings.
- Construct validity may be established by determining whether the test discriminates between students who know the content and those who do not.

SET CUT SCORES

Introduction

A cut score, or standard, is the test score a student must achieve in order to "pass" the test. Cut scores must be set in a manner which is defensible and fair and by persons qualified to make such decisions. This issue is of major importance, and will be briefly discussed here. Those qualified to make judgments about cut scores will be those most familiar with the content area. For example, auto mechanics teachers and auto mechanics themselves should be used to make judgments about the auto mechanics items. Teachers and incumbent workers who serve to determine what should be included on the test may also be asked to provide judgments regarding the appropriate cut score for that test.

Categories of Cut Score Methodologies

There are two categories of methodologies used for setting cut scores. The first is based on judgments about items, while the second is based on judgements about individuals. Three common methodologies will be briefly described below. Individuals desiring further information may refer to Livingston & Zieky (1982).

Methods Based on Judgments About Items: The Angoff Method

One method for setting cut scores based on judgments about items was proposed by William Angoff (1971). In this method, the judge considers each item as a whole and makes a judgment about the probability that a borderline or minimally qualified student would answer the item correctly (or would perform the item adequately). In other words, the judges indicate the percent of minimally competent students they think would pass the item. The probability (or percentage) ranges from .00 to 1.00. A probability of .00 means that the borderline student would have no chance of performing the item correctly, while a 1.00 means that borderline students would definitely be able to perform the item correctly. The higher this probability, the more likely the judge thinks it is that a borderline student will be able to pass that item.

Judges are asked to make judgments about a few items, then these judgments are discussed. Items where the probabilities assigned by the judges are more than 15 to 20 percentage points different are discussed and a consensus reached. Judges then finish the task and discuss all items after all judgments have been made. To calculate the cut score, simply find the mean (i.e., average) probability and multiply by the number of items (or points).

Example: Angoff Equally Weighted Items

Assume you have four judges and five items and the following probabilities suggested as the percentage of minimally competent students passing the item.

	Judges			
	Judge A	Judge B	Judge C	Judge D
Item				
1	.86	.34	.77	.63
2	.91	.47	.88	.80
3	.75	.76	.72	.70
4	.65	.67	.71	.58
5	.34	.27	.65	.44

Notice that items 1, 2, and 5 should be discussed by the judges since there is quite a large disparity among these percentages. After discussion, assume the following probabilities (percentages) are obtained:

	Judges			
	Judge A	Judge B	Judge C	Judge D
Item				
1	.86	.84	.77	.83
2	.91	.87	.88	.80
3	.75	.76	.72	.70
4	.65	.67	.71	.58
5	.34	.27	.35	.34

Note the greater consistency of scores after they have been discussed. Now, a cut score can be calculated as follows.

	Judges			
	Judge A	Judge B	Judge C	Judge D
Item				
1	.86	.84	.77	.83
2	.91	.87	.88	.80
3	.75	.76	.72	.70
4	.65	.67	.71	.58
5	.34	.27	.35	.34
Sum	3.51	3.41	3.43	3.25
Mean	.70	.68	.69	.65

Sum of the Mean Percentages = 2.72
Grand Mean = .68

Thus, the cut score would be 68%. Let's assume there are 55 points on the exam, and the final cut score would be .68 X 55 = 37.4 or 37 points.

Example: Angoff Differentially Weighted Items

The above example assumes that all items are equally weighted. However, since each item is considered separately in the Angoff method, weightings attached to each item may also be attached to the percentages in the same manner as they are applied to the items. Remember, however, to consider this when calculating the means, as the number of "items" will change to reflect the differential weights. Let's look at an example.

		Judges			
		Judge A	Judge B	Judge C	Judge D
Item	Points				
1	2	.86 (1.72)	.84 (1.68)	.77 (1.54)	.83 (1.66)
2	3	.91 (2.73)	.87 (2.61)	.88 (2.64)	.80 (2.40)
3	4	.75 (3.00)	.76 (3.04)	.72 (2.88)	.70 (2.80)
4	1	.65 (0.65)	.67 (0.67)	.71 (0.71)	.58 (0.58)
5	2	.34 (0.68)	.27 (0.54)	.35 (0.70)	.34 (0.68)
Sum	12	3.51 (8.75)	3.41 (8.54)	3.43 (8.47)	3.25 (8.12)
Mean		.70 (0.73)	.68 (0.71)	.69 (0.71)	.65 (0.68)

Notice that the number of points for each item has been multiplied by the probability assigned by each judge. For example, item 3 is weighted 4 points and Judge B assigned a probability of .76. Thus, the total for item 4 for Judge B is 4 X .76 or 3.04. The products of the weights and the probability are indicated in parentheses in the above example. The cut score is set in the same way as it is for unweighted items, that is, you add up all the average probabilities for each judge and divide by the number of points. Notice, however, that now the number of points is 12, not five! These mean probabilities are

then added and divided by the number of judges (4). Thus, to calculate the cut score for this example, you add .73, .71, .71, and .68 (the sum is 2.83) and divide by the number of judges (4). The cut score is .71 or .71 X 12 = 8.52, which rounds up to 9 points (out of 12). In this manner, items having more weight on the test also have more weight in determining the cut score; in fact, the weightings are identical.

Modifications of the Angoff Method

There are a number of modifications of the Angoff method, most involving providing judges with some sort of student performance data. These modifications generally involve following the process above first without student performance data, and then again with student performance data provided. The type of student performance data usually supplied is the mean performance of students, the p value (difficulty index) for each item, the standard deviation, etc. The p value, or difficulty index, is an index which indicates the difficulty of the item, just as the name implies. The higher the p value, the easier the item and the lower the p value, the harder the item. This seems to be backwards to many people, but if you understand how it is calculated, it makes more sense. In order to calculate the p value, simply divide the number of people who got the item correct (or passed it) by the total number that took the item. Thus, if 35 students took an item and 17 of them got it correct (passed it), the p value is 17/35, or .49. This means that 49% of the students who took the item got it right. The p value will range from .00 to 1.00. The ideal p value will vary depending on the nature of the test. For some tests (e.g., competitive tests for scholarships), it is desirable for the p value to be quite low (in other words, have very difficult items) so that only the top students pass the test. For other tests (such as the minimum competency tests like the Gold Seal Endorsement), it is desirable for the p value to be a bit higher, so that any student who is minimally qualified or more will pass the item.

Methods Based on Judgments About Items: The Ebel Method

Ebel's (1972) method requires that judges make two kinds of judgments about each item--a judgment of the item's difficulty and a judgment about the item's relevance or importance. Ebel suggests three levels of difficulty--easy, medium, and hard; and four levels of relevance--essential, important, acceptable, and questionable. Judges are asked to classify all items into one of the twelve categories (e.g., essential and easy, important and hard). A table similar to the following will result.

Relevance	Difficulty Easy	Medium	Hard
	Items	Items	Items
Essential	1, 7, 12	6, 13, 9	3, 8
Important	2, 4, 15	11, 14	10
Acceptable	5		
Questionable	17		

The judge must then make judgments about the performance of minimally qualified students for each of the cells in the table which contain items. The judge must ask her or himself, "If a borderline student were to take a lot of items like these, what percentage of the items would he or she answer correctly?" The judgments may then be added to the table as in the example below.

Example: Ebel

Assume the following table is generated.

Relevance	Difficulty Easy	Medium	Hard
	Items	Items	Items
Essential	1, 7, 12	6, 13, 9	3, 8
	.98*	.85	.75
Important	2, 4, 15	11, 14	10
	.91	.81	.65
Acceptable	5		
	.95		
Questionable	17		
	.91		

* These are average percentages of four judges

As with the Angoff method, judges should discuss preliminary judgments on a few items and then continue making judgments after reaching a compromise on these items. After all judgments have been made, judges should discuss any judgments which are more than 15 or 20 percentage points disparate. To obtain the cut score, for each category, multiply the judged percentage correct by the number of questions in that category. This will provide the expected score for that cell. Add the expected score for all cells and obtain the expected score for the entire test. An example is shown on the next page. For the purposes of the example, assume that there are 16 points on the test.

Note that the Ebel method assumes that all items are equally weighted. If the items are not equally weighted, some sort of weighting procedure similar to the one applied on the test must be made. This makes this method much more difficult and unwieldy to perform because items are not individually assigned a percentage. Thus, the Ebel method is most applicable to situations where items are equally weighted. In addition, it should be noted that items which fall into the "Questionable"

category in terms of relevance should probably be dropped, at least from the performance test. Performance testing is such an expensive enterprise that only essential and important tasks should be included.

Category	Percent Correct	No. of items	Expected Score
Essential			
Easy	.98	3	.98 X 3 = 2.94
Medium	.85	3	.85 X 3 = 2.55
Difficult	.75	2	.75 X 2 = 1.50
Important			
Easy	.91	3	.91 X 3 = 2.73
Medium	.81	2	.81 X 2 = 1.62
Difficult	.65	1	.65 X 1 = 0.65
Acceptable			
Easy	.95	1	.95 X 1 = 0.95
Medium	--	-	--
Difficult	--	-	--
Questionable			
Easy	.99	1	.99 X 1 = 0.99
Medium	--	-	--
Difficult	--	-	--
SUM			13.93

which rounds up to 14.

Methods Based on Judgments About Individuals

One method based on judgments about individuals, the Borderline Groups Method, is described here. This method proposes that the cut score should be the score which would be expected from a student who is borderline. Instead of asking judges to make judgments about how a borderline student would perform, the judges actually identify borderline students. The students are then given the test and the cut score is then the median score (50th percentile) of the borderline group. The median score is the score which divides the group exactly in half; that is, half of the persons are above and half are below the median score.

Example

Assuming there are 100 points in the test, then the median score is:

Score	Rank Order of Score (from lowest to highest)
78	49
75	53
53	62
49	66
62	75
66	78

The median score is equal to 64 ([62 + 66]/2). If there is an even number of scores (as in the above example), the median is computed by rank ordering the scores, taking the two middle scores (62 and 66), adding them, and then dividing by 2. If there is an odd number of scores, rank order the scores and the median will be the score which is in the exact middle. For example:

Score	Rank Order of Score (from lowest to highest)
78	49
75	53
53	62
49	66
62	75
66	78
80	80

In this case, the median score is 66.

Other Methods

Note that there are two other popular methods: the Nedelsky (1954) method and the Contrasting Groups method. The Nedelsky method involves judgments about items and may only be applied to multiple choice items. The Contrasting Groups method, as the name implies, requires judgments about individuals. However, it can become quite complex. Readers interested in this method may refer to Livingston & Zieky (1982).

Summary

The following steps may be taken to determine the cut score.

- Determine the appropriate cut score methodology.
- Determine the persons who may serve as judges.
- Train judges in the cut score methodology selected.
- Judges set cut score.

OTHER ISSUES

Performance tests can be prone to low reliability and high errors. There are several ways in which errors may be reduced. These are:

1. standardize the testing conditions,
2. train raters,
3. use more than one rater,
4. ask raters to record their observations as soon as possible after making them, and
5. give raters examples of adequate and inadequate examples of performance. Borderline examples should also be included.

Once the test has been developed and used, it must be periodically reviewed to update and improve it to reflect changes in technology, etc. In addition, as the test is administered over time, data should be accumulated on means, standard deviations, standard errors, reliability, and validity, as well as demographic factors. Such demographic factors should include sex, race, and, perhaps, age. Data may then be analyzed to determine if there are any score differences in performance between the different groups. Just like a written test, a performance test should not discriminate on the basis of race, sex, handicap, or membership in a special population. In addition, any test should be used only for the purpose for which it was intended. Use for other purposes may have adverse legal consequences.

Summary

Some further issues and considerations in performance testing include the following.

- Reduce errors by
  a. standardizing the testing conditions,
  b. training the raters,
  c. using more than one rater,
  d. asking raters to record their observations as soon as possible after making them, and
  e. giving raters examples of adequate and inadequate examples performance. Borderline examples should also be included.
- Periodically review the test and update it.
- Determine if there are differences in performance between different groups. The test should not discriminate between different groups of students.
- Use the test only for the purpose for which it was intended.

AN EXAMPLE

Let's suppose, for our example, that you work for Utopia County and it is your responsibility to develop and oversee the administration of the Gold Seal Endorsement program. Part of that responsibility involves the development of a performance test for the area of Automotive Mechanics. It is the purpose of this example to demonstrate each step in the process of development of a performance test.

DETERMINE WHAT SHOULD BE TESTED (Manual reference pp. 2-5)

First, you must decide what should be included on the performance test. When choosing tasks for inclusion on a performance test, the following steps may be taken:

- Involve content experts in selecting tasks to be included on the performance test.
- Select a representative sample of critical tasks for inclusion on the performance test.
- Consider the time available for testing and the time required for the tasks under consideration.
- Consider resources available and other assessment methodologies.
- Break down complex tasks into their basic components.
- Avoid testing personal values or attitudes, if possible.
- Select tasks appropriate for entry level job performance.

Suppose that you have a degree in, let's say, Educational Administration, and know little or nothing about Automotive Mechanics beyond "The car is making funny noises", "I need to get gas", and "The tire is flat." Obviously, you need help in deciding what should be on the test. You call in a group of Automotive Mechanics teachers from across Utopia County to help you make your decision. Let's assume that Utopia County is moderate in size and serves both urban and rural students. You decide that three teachers will be sufficient. You call a teacher from Nirvana High School, a suburban high school; one teacher from Garden of Eden High School, a rural high school; and one teacher from Ignorance is Bliss High School, an urban high school. These teachers, along with a mechanic from Shirley's Shangri-La Garage will help you decide which tasks to include on the test.

These four committee members decide that the following tasks form a representative sample of performance which could reasonably be expected from a student graduating from one of Utopia's high schools in Automotive Mechanics:

- Reface valves and seats.
- Record measurements and compare with specifications in service manual.
- Inspect front suspension and steering systems.
- Perform standard tests on engine electrical system.
- Check two-wheel alignment.
- Inspect and evaluate front disc brakes.
- Use radiator and antifreeze tester to evaluate and pressure-test the system.
- Pressure test the air conditioning system for leaks.

Note that many of these tasks cover more than one student performance standard. For example, perform standard tests on engine electrical system covers the following eight student performance standards:

- Use and apply basic electrical and electronic test equipment and meters.
- Perform continuity tests.
- Measure voltage drop, current flow and resistance in a circuit or component with a multimeter.
- Analyze cranking system malfunctions.
- Analyze charging system malfunctions.
- Service and test batteries.
- Test, remove and replace fuses and circuit breakers.
- Inspect and test primary circuits.

Let's assume that students in Utopia County attend class in two-hour blocks of time and that no more time can be given for testing purposes. Thus, the performance test must be conducted over a number of days, rather than all at once. All tasks included on the committee's list of representative tasks can be performed within an hour. Let's assume that the committee also determines that five hours of testing will be sufficient and that the above list requires more than five hours. The committee then notes that one of the items, "recording measurements and comparing them to specifications in a service manual", can be easily tested using paper and pencil methodologies. Thus, this task is deleted from the performance test. The committee feels that the remaining tasks can easily be done in five hours. They therefore set the total testing time at five hours.

Given time and space requirements, only one of the above tasks, "check two-wheel alignment", is discussed further in this manual. [As an aside, apologies are extended to Automotive Mechanics teachers who may note inaccuracies in this example (for example rating one task as more important than another when this does not accurately reflect the actual importance of the tasks). This example is provided mainly for illustrative purposes, and the accuracy of some of the suppositions may be questionable.] The task "check two-wheel alignment" can be broken down into the following component steps:

Measure and correct ride height
Read and correct the caster
Read and correct the camber
Read and correct the toe in
Center the steering wheel

DEVELOP TEST BLUEPRINT (Manual reference pp. 6-12)

Now that the tasks to be included on the performance test have been selected and broken down into their component parts, a test blueprint may be developed. It will be remembered that the blueprint contains information about test length, scoring methodologies, weighting of different skills, when to test, the conditions under which testing is to occur, who can serve as a rater, test security, etc.

As mentioned previously test length is related closely to what is to be tested. In our example, the skills to be tested determined to a great extent the test length. As discussed on the previous page, the committee set total testing time at five hours.

Now, issues of scoring must be addressed. First, must every skill be passed? In our example, there are five skills which the student must demonstrate in order to check two-wheel alignment. It must be decided by content experts whether or not the student must pass each one of these five skills or whether one or more of the skills may be omitted or not passed. For example, suppose our committee of content experts decides that the student does not need to pass each skill in order to receive credit for the entire item "checking two-wheel alignment". If the student omits one of the component skills or performs it incorrectly, the item may still be marked as passed.

The next decision involves whether there are certain skills the students must pass in order to receive credit. In our example, say the committee decides that it is essential that the student be able to read and correct the caster. If this task is not passed, the student cannot pass the item, regardless of how well they performed the other tasks (such as reading and correcting the camber).

Now the committee must decide whether or not every skill will receive a score or whether an overall score will be assigned for the item. Further, they must decide if every item will receive a score or whether an overall score will be assigned for the performance test. Our committee decides that each item will receive a score, rather than each skill. The process by which the committee comes to this decision is quite subjective and often involves lively discussion among members.

A related issue is whether analytic or holistic scoring is to be used. In analytic scoring, each criterion or component of a task is evaluated individually and assigned a score. In holistic scoring, all criteria are considered simultaneously and one

overall score is assigned. Even with holistic scoring, some criteria can be given more weight than others. Holistic scoring is less time consuming and less expensive but doesn't provide as much information. Analytic scoring can provide diagnostic information for remediation if that is appropriate. Our committee decides that holistic scoring is more appropriate since information for remediation is not considered part of this assessment and it is somewhat easier and less expensive than analytic scoring.

Next, the committee must decide whether a rating scale or checklist is to be developed for each item. Often, this decision will be a matter of personal preference and/or ease of development and use. Our committee decides that since an overall score will be assigned each item and since checklists are a little easier to develop and use, the checklist is to be used for all items.

Finally, the issue of process vs. product must be decided. In some cases, as in our example, this decision will be made for us by the task. In our example, the final "product" is the degree with which the student has centered the steering wheel. However, suppose our committee has decided that the process should also be measured. In some examples, the process may be wholly or in part unobservable (for example, playing the piano or composing a letter of resignation) and the product must be measured. In still other examples, there may be no product (e.g.,"Read a story from a picture book to a groups of eight to ten preschool children."). For the entire test, it is best to use some items which may be scored using process evaluations and some items which may be scored using product evaluations if possible.

The next issue in the blueprint will concern whether or not items should be differentially weighted. Again, this will be a subjective decision on the part of committee members. In our example, several of the tasks measure more student performance standards and are more complex than others. In addition, some of the tasks are performed more frequently than others on the job. Thus, our committee decides that these tasks should be given more weight than the less complex and less frequently occurring tasks. A hypothetical example of the weights assigned to the tasks might be:

- Reface valves and seats -- 5 points.
- Inspect front suspension and steering systems -- 5 points.
- Perform standard tests on engine electrical system -- 25 points.
- Check two-wheel alignment -- 5 points.
- Inspect and evaluate front disc brakes -- 15 points.
- Use radiator and antifreeze tester to evaluate and pressure-test the system -- 10 points.
- Pressure test the air conditioning system for leaks -- 25 points.

Notice that the sum of the points need not add up to 100.

The issue of when to test will be determined by the purpose of the test and the time available for testing (for example two hour blocks). Our committee decides that this is an end of program test and should be administered at the end of the program in a structured setting.

The setting under which students are administered the test must be as nearly alike for each student as possible so that it is fair to all. For example, if one student received a piece of equipment which was not calibrated properly while another student's equipment was properly calibrated, it would not be fair to compare the performance of these two students. Thus, the conditions under which testing will occur should be clearly specified. For our check two-wheel alignment example, the committee decides that the vehicle must be on a rack and elevated in the shop. In addition, the ride height, castor, camber, toe in, and steering wheel should be out of alignment to a similar degree for all students. The student should be given 60 minutes to read directions and perform the item.

The issue of who may serve as raters is an important one. The raters should be familiar with the tasks the student is to perform. Thus, our committee decides that Automotive Mechanics teachers will serve as raters. Each student will be assessed by two Automotive Mechanics teachers. The student's classroom teacher should not be involved in serving as a rater for that particular student.

The committee then addresses the issue of test security. In performance testing, test security is less important than in paper and pencil tests. Thus, our committee decides that no test security is necessary.

DEVELOP THE TEST (Manual reference pp. 13-18)

Once the blueprint is developed and decisions made, the next step is to develop the test. This is now easily accomplished, as all difficult decisions have been made during the blueprint process. The first step is to develop instructions for the raters. In our example, instructions to raters might read something like the following:

> The student is to inspect and indicate adjustments needed to the wheel system to restore the vehicle to factory specifications. You are to use the scoring checklist to rate the student's performance as either Acceptable or Unacceptable.
>
> Give the students the Directions to the Student and the Scoring Checklist. Timing of the activity should be as follows:
>
> 1. Give the student five minutes to read the Directions to the Student and the Scoring Checklist. Ask the student if there are

any questions. Answer the student's questions.

2. Give the student up to 60 minutes to complete the task.

Notice that the directions to the raters do not include directions on setting up the item. This will normally be done in advance by someone else, usually a test administrator, although a teacher could do it. Separate directions for setting up the area should be included. These directions may read something like the following:

> Set up all necessary equipment and supplies in the area where the testing will take place (see listed necessary equipment and supplies).
>
> All vehicles must be on the rack and raised to the appropriate height for the student. In addition, all the vehicles' wheel systems should be equally out of alignment.
>
> Provide the raters with the performance objective, the list of necessary equipment and supplies, Directions to the Raters, Directions to the Students, and the Scoring Checklist. Provide extra copies of the Directions to the Students and Scoring Checklist so that all students to be tested may be provided with a copy.

Now, Directions to the student must be developed. For our example, these directions might read something like the following:

> Your task is to inspect the vehicle on the rack and make any needed adjustments to the wheel system to restore the vehicle to factory specifications.
>
> You have five minutes to read these directions and the Scoring Checklist. The Scoring Checklist indicates how you will be scored on this item.
>
> You have up to sixty minutes to complete the following tasks:
>
> Measure and correct the ride height.
> Read and correct the caster.
> Read and correct the camber.
> Read and correct the toe in.
> Center the steering wheel.
>
> If you have any questions, please ask now.

Now, all equipment, materials, and supplies necessary for performance of the task must be specified. For our example, the following equipment, supplies, and materials are needed:

Equipment	Supplies	Materials
-safety glasses	-paper towels	-manufacturer's specifi-
-basic hand tools	-fender covers	cations
-alignment rack	-alignment shims	-shop manuals
-alignment equipment		
-specialized alignment tools		

Now it is time to develop the rating scales and/or checklists to be used in scoring. While it was decided that a checklist was to be used in our example, a sample of each type of scoring document is shown on the following two pages.

PERFORMANCE CHECKLIST

Directions: Observe the student as he or she performs the task. As the student performs each task, place a check (✓) or X in the correct column (Yes or No).

	Yes	No
1. Measure and correct the ride height to within 1/2 inch of manufacture's specification.	_____	_____
2. Read and correct the caster to within 1 degree of manufacturer's specification.	_____	_____
3. Read and correct the camber to within 1/2 degree of manufacturer's specification.	_____	_____
4. Read and correct the toe in to within 1/4 degree of manufacturer's specification.	_____	_____
5. Center the steering wheel to within 2 degrees of manufacturer's specification.	_____	_____

COMMENTS:

PERFORMANCE RATING SCALE

Directions: Observe the student as he or she performs the task. As the student performs each task, circle the number on the rating scale which indicates the accuracy with which the task is completed.

Task 1: Read and correct the ride height.

Error of more than 1.5" from manufacturer's spec.	Error of 1.1" to 1.5" from manufacturer's spec.	Error of .5" to 1" from manufacturer's spec.	Error of less than .5" from manufacturer's spec.
1	2	3	4

Task 2: Read and correct the caster.

Error of 4 degrees or more from manufacturer's spec.	Error of 3 degrees from manufacturer's spec.	Error of 2 degrees from manufacturer's spec.	Error of 1 degree or less from manufacturer's spec.
1	2	3	4

Task 3: Read and correct the camber.

Error of 3 degrees or more from manufacturer's spec.	Error of 2 degrees from manufacturer's spec.	Error of 1 degrees from manufacturer's spec.	Error of 1/2 degree or less from manufacturer's spec.
1	2	3	4

Task 4: Read and correct the toe-in.

Error of 2 degrees or more from manufacturer's spec.	Error of 1 degrees from manufacturer's spec.	Error of 1/2 degree from manufacturer's spec.	Error of 1/4 degree or less from manufacturer's spec.
1	2	3	4

Task 5: Center the steering wheel.

Error of more than 10 degrees from center	Error of 6 to 10 degrees from center	Error of 2 to 5 degrees from center	Error of less than 2 degrees from center
1	2	3	4

COMMENTS:

TRAIN THE RATERS (Manual reference pp. 18-19)

Now that the content has been decided upon and the blueprint and test have been developed, raters can be trained. Raters should be familiarized with the tasks to be performed and the criteria to which tasks should be performed. The training for the raters should also involve how to use the checklists and/or rating scales which have been developed. In addition, raters should be made aware of common rater biases, such as the halo effect and leniency or stringency effects. Raters can then guard against allowing such biases to influence their ratings. Indices of intra-rater reliability should be gathered during the training process.

PILOT TEST THE INSTRUMENT (Manual reference p. 19)

Pilot testing the instrument is an important process which should not be overlooked. The pilot test will indicate errors such as materials and/or equipment that may have been omitted from the list, that the rating scale or checklist is difficult and unwieldy to use, that the time limit set is either too stringent or too lenient, that the directions to the student, raters, and/or test administrators are unclear, etc. Pilot testing need not be done on a large number of students for the purposes of this performance test. It is, however, important to include students who will represent an entire range of possible performance.

DETERMINE RELIABILITY AND VALIDITY (Manual reference pp. 19-24)

Only the aspects of reliability and validity remain to be settled before the testing process is ready to commence. Reliability should be determined by a test-retest method, if possible, as well as inter-rater reliability estimates. If using a test-retest method, be sure that the student is competent in the tasks at both testing periods and that enough time has elapsed so that memory will have minimal effects. For the purposes of our example, we will assume that the test is administered at three month intervals to establish test-retest reliability. Also for our example, assume that inter-rater reliability is determined by obtaining the correlation between the ratings given by the two raters. During pilot testing, the correlation coefficient was found to be .87. This indicates very good agreement between the two raters. Reliability is an on-going process and inter-rater reliability should be checked for each administration.

The content validity of the test was probably assured during the developmental phase by involving Automotive Mechanics teachers and a mechanic in the process of selecting items for inclusion on the performance test. However, periodic review of the performance test for content validity is necessary.

For our example, let's suppose some of the students work part-time in Shirley's Shangri-La Garage. Shirley has rated these

students as she does all her employees. Thus, we can establish criterion-related validity by correlating the scores the students obtain on the performance test with the supervisory ratings Shirley has given the students.

We can establish construct validity by testing students across the range of possible proficiency. If a student who is known to be proficient at a task and one who is known to be much less proficient at the task are given the test, the two should receive scores which would suggest their relative ranking. The student known to be proficient should receive a relatively high score, while the student who is not proficient should receive a relatively low score.

SET CUT SCORE (Manual reference pp. 24-30)

Next, a cut score should be set. Let's assume that an Angoff approach was utilized, set by the teachers and auto mechanic who originally served on our committee. For our example, assume that the cut score was set to be 87%.

SUMMARY

Now that all issues have been decided and the test developed, we would have something that looked like the following pages. This outline summarizes the process of developing the Automotive Mechanics performance test.

A. Content to be included on test
   - Reface valves and seats.
   - Inspect front suspension and steering systems.
   - Perform standard tests on engine electrical system.
   - Check two-wheel alignment.
   - Inspect and evaluate front disc brakes.
   - Use radiator and antifreeze tester to evaluate and pressure-test the system.
   - Pressure test the air conditioning system for leaks.

B. Test Blueprint
   1. Test Length
      Determined by tasks included on test.
   2. Scoring
      a. Must student pass every skill? No
      b. Are there certain tasks the student must pass? Yes, see specific items
      c. Analytic vs. holistic scoring? Holistic

3. Weighting? Yes

Skill	Points
--Reface valves and seats	5
--Inspect front suspension and steering systems points.	5
--Perform standard tests on engine electrical system.	25
--Check two-wheel alignment.	5
--Inspect and evaluate front disc brakes.	15
--Use radiator and antifreeze tester to evaluate and pressure-test the system.	10
--Pressure test the air conditioning system for leaks.	25

4. When to test? End of program
5. Conditions under which testing occurs. Standardized, see Directions to Administrator for specific conditions for each item.
6. Who may serve as raters? Two teachers per student
7. Test security? None required.

C. Develop the Test--see final example

D. Train raters

E. Pilot test the instrument

F. Determine reliability
   1. test/retest--readminister test after 3 months.
   2. inter-rater reliability determined by correlations between ratings of two observers.

G. Determine validity
   1. Content validity--established during development by involving content experts in decision regarding performances to be included on the test.
   2. Criterion related validity--established by correlating scores on performance test with supervisory ratings.
   3. Construct validity--established by looking at scores obtained by students with different known levels of competence on items.

H. Determine cut score
   1. Used Angoff procedure
   2. Three teachers and one auto mechanic served on committee
   3. Cut score is 87%.

THE ACTUAL ITEM

Program: Automotive Mechanics

Performance Objective: Inspect vehicle on rack and make adjustments needed to wheel system necessary to restore to specifications.

Requirements:

Equipment	Supplies	Materials
-safety glasses	-paper towels	-manufacturer's specifications
-basic hand tools	-fender covers	
-alignment rack	-alignment shims	-shop manuals
-alignment equipment		
-specialized alignment tools		

Directions to Administrator:

Set up all necessary equipment and supplies in the area where testing will take place (see listed necessary equipment and supplies).

All vehicles must be on the rack and raised to the appropriate height for the student. In addition, all the vehicles' wheel systems should be equally out of alignment.

Provide the raters with the performance objective, the list of necessary equipment and supplies, Directions to the Raters, Directions to the Students, and the Scoring Checklist. Provide extra copies of the Directions to the Students and Scoring Checklist so that all students to be tested may be provided with a copy.

Directions to the Rater:

The student is to inspect and make adjustments needed to the wheel system to restore the vehicle to factory specifications. You are to use the scoring checklist to rate the student's performance as either Acceptable or Unacceptable.

Give the students the Directions to the Student and the Scoring Checklist. Timing of the activity should be as follows:

1. Give the student five minutes to read the Directions to the Student and the Scoring Checklist. Ask the student if there are any questions. Answer the student's questions.
2. Give the student up to 60 minutes to complete the task.

Directions to the Student:

Your task is to inspect the vehicle on the rack and make any needed adjustments to the wheel system to restore the vehicle to factory specifications.

You have five minutes to read these directions and the Scoring Checklist. The Scoring Checklist indicates how you will be scored on this item.

You have up to sixty minutes to complete the following tasks:

Measure and correct the ride height.
Read and correct the caster.
Read and correct the camber.
Read and correct the toe in.
Center the steering wheel.

If you have any questions, please ask now.

PERFORMANCE CHECKLIST

Directions: Observe the student as he or she performs the task. As the student performs each task, place a check (✓) or X in the correct column (Yes or No).

	Yes	No
1. Measure and correct the ride height to within 1/2 inch of manufacture's specification.	_____	_____
2. Read and correct the caster to within 1 degree of manufacturer's specification.	_____	_____
3. Read and correct the camber to within 1/2 degree of manufacturer's specification.	_____	_____
4. Read and correct the toe in to within 1/4 degree of manufacturer's specification.	_____	_____
5. Center the steering wheel to within 2 degrees of manufacturer's specification.	_____	_____

COMMENTS:

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, D.C.: American Psychological Association.

Angoff, W.H. (1971). Scales, Norms, and Equivalent Scores. In R.L. Thorndike (ed.). Educational Measurement. Washington, D.C.: American Council on Education.

Cohen, J. (1965). Some statistical issues in psychological research. In B.B. Woman (ed.), Handbook of clinical psychology. New York: McGraw-Hill.

Ebel, R.L. (1972). Essentials of Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall.

Klein, R. (1980). Some Selected Technical Issues Related to Performance Testing. In Performance testing: Issues facing vocational education. J. E. Spirer (ed). Research and Development Series No. 190. Columbus, OH: The National Center for Research in Vocational Education, The Ohio State University.

Livingston, S.A. (1980). Comments on the Technical Issues in Performance Testing. In Performance testing: Issues facing vocational education. J. E. Spirer (ed). Research and Development Series No. 190. Columbus, OH: The National Center for Research in Vocational Education, The Ohio State University.

Livingston, S.A. & Zieky, M.J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests. Monterey, CA: CTB/McGraw Hill.

Messick, S. (1989). Validity. In Educational Measurement (3rd ed). R. L. Linn (ed). New York: American Council on Education.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14(1), 3-19.

Priestley, M. (1982). Performance assessment in education and training: Alternative techniques. Englewood Cliffs, NJ: Educational Technology Publications.

Simon, J.L. (1978). Basic Research Methods in Social Science: The Art of Empirical Investigation. (2nd ed.). New York: Random House, p. 273.

OTHER READINGS OF INTEREST

Berk, R.A. (ed). (1986). Performance assessment: Methods & applications. Baltimore: The Johns Hopkins University Press.

Fitzpatrick, R. & Morrison, E.J. (1971). Performance and Product Evaluation. In Educational Measurement (2nd ed). R.L. Thorndike (ed). Washington, D.C.: American Council on Education.

Sanders, J.R., & Sachse, T.P. (1977). Applied performance testing in the classroom. Journal of Research and Development in Education, 10(3), 92-104.

Stiggins, R.J. (1987). Design and Development of Performance Assessments. Educational Measurement: Issues and Practice. 6(3), 33-42.

Traub, R.E. & Rowley, G.L. (1991). Understanding reliability. Educational Measurement: Issues and Practice, 10, 37-45.

APPENDIX

A correlation coefficient, symbolized by r_{xy} indicates the strength of relationship between two variables. For test-retest reliability, if r_{xy} is high and positive, it means that a student with a high score on the first administration of the test is likely to score high on the retest. This is what we would want to observe. One formula for computing the correlation coefficient (more formally called the Pearson Product Moment Correlation Coefficient) is:

$$r_{xy} = \frac{N \Sigma XY - (\Sigma X)(\Sigma Y)}{[N \Sigma X^2 - (\Sigma X)^2][N \Sigma Y^2 - (\Sigma Y)^2]}$$

Where: Σ = the Greek letter capital sigma. This means that you are to add the term following it. For example, ΣX means to add all the X scores.
N = the number of items
X = score on the first administration
Y = score on the second administration

For example, suppose we have the following scores from two administrations of the test.

Item	X	Y
1	10	9
2	7	8
3	6	5
4	3	2
5	5	5
6	4	6
7	8	6
8	9	10
9	7	5
10	1	3

In order to compute the correlation coefficient, we can follow the steps given below.

1) Add up the X scores (ΣX)
2) Add up the Y scores (ΣY)
3) Square each of these sums [$(\Sigma X)^2$ and (ΣY^2)]
4) Square the X scores and add them up (ΣX^2)
5) Square the Y scores and add them up (ΣY^2)
6) Multiply each X and Y score and add up the products (ΣXY)
7) Substitute each of these terms into the appropriate place in the formula and carry out the computations.

In our example, we have the following.

Item	X	Y	X^2	Y^2	XY
1	10	9	100	81	90
2	7	8	49	64	56
3	6	5	36	25	30
4	3	2	9	4	6
5	5	5	25	25	25
6	4	6	16	36	24
7	8	6	64	36	48
8	9	10	81	100	90
9	7	5	49	25	35
10	1	3	1	9	3
= SUM	60	59	430	405	407

Therefore, $\Sigma X = 60$; $\Sigma Y = 59$; $\Sigma X^2 = 430$; $\Sigma Y^2 = 405$; and $\Sigma XY = 407$.

$$\text{Thus } r_{xy} = \frac{10(407) - 60(59)}{[10(430)-(60)^2][10(405)-(59)^2]}$$

$$= \frac{530}{631.11} = .84$$

Thus, students who scored high on the first administration of the test also scored high on the second administration of the test. Looking at the raw data, it can be seen that this is the case.

There are several "canned" computer programs which calculate the Pearson Product Moment (PPM) Correlation Coefficient. If the program calculates more than one type of correlation coefficient, be sure to indicate that the PPM correlation coefficient is to be used to calculate test-retest reliability. You can also use many hand calculators, particularly those with statistical functions which have special keys for a variety of common statistical calculations. A programmable calculator may also be used. Once it is programmed and the program stored into memory, it can be used again without reprogramming.

To calculate inter-rater reliability, the same procedure as above would be used <u>if</u> the raters assign some sort of total score to each item included on the performance test or if a rating scale is used. However, if a checklist is used, a slightly different correlation coefficient must be used. This is called the phi coefficient. Symbolized ϕ, this correlation coefficient is used when the two scores to be correlated are dichotomous (are either "yes/no", "pass/fail", etc.). For the following examples, we will assume that "1" indicates that the student received a "yes" or "pass" while a "0" indicates that the student received a "no" or "fail."

To calculate ϕ, one sets up a fourfold table as follows:

		Rater 2	
		1	0
Rater 1	0	a	b
	1	c	d

Each cell in the table is labeled. Thus, for example, a is the cell where rater 1 gave the student a "0" and rater 2 scored the student as "1." Phi is then calculated as follows:

$$\phi = \frac{bc - ad}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$$

Suppose we have the following data.

Item	Rater 1	Rater 2
1	1	1
2	0	1
3	1	1
4	1	1
5	0	0

Fitting this data into the table, we obtain:

		Rater 2	
		1	0
Rater 1	0	1	1
	1	3	0

This table merely tells us that both Rater 1 and Rater 2 gave 3 of the students "pass"; Rater 1 and Rater 2 both rated the same student as "fail"; and that Rater 2 rated one student "pass" while Rater 1 rated that student "fail".

Thus, ϕ becomes:

$$\phi = \frac{1(3) - 1(0)}{\sqrt{(1+1)(1+3)(3+0)(1+0)}}$$

$$\phi = 3/4.9 = .612$$

A second method for calculating inter-rater reliability involves dividing the number of agreements by the total number of ratings. This method would be used for checklists with dichotomous data.

For the data above, we would calculate the inter-rater reliability in the following manner:

Item	Rater 1	Rater 2	
1	1	1	(both raters agree on item 1)
2	0	1	(raters disagree on item 2)
3	1	1	(both raters agree on item 3)
4	1	1	(both raters agree on item 4)
5	0	0	(both raters agree on item 5)

Number of agreements = 4
Total number of ratings = 5

Agreement = 4/5 = .80

Notice that this gives you a much higher estimate than does ∅. This is because ∅ is a correlation coefficient while this method yields a percentage of rater agreement.

The third method of computing inter-rater reliability is Cohen's kappa (k). The formula for k is:

$$k = \frac{P_o - P_c}{1 - P_c}$$

Where: P_o = the proportion of agreements between raters. P_c = the proportion of agreements expected by chance. To compute P_c, multiply the number of "1's" (or "yes" or "pass" scores) for Rater 1 by the number of "1's" for Rater 2; multiply the number of "0's" (or "no" or "fail" scores) for Rater 1 by the number of "0's" for Rater 2; then add the two products and divide by the total number of observations squared.

Using the above data, we have,

Item	Rater 1	Rater 2
1	1	1
2	0	1
3	1	1
4	1	1
5	0	0

The two raters agree on 4 out of 5 of the items, so P_o = 4/5 = .80. Rater 1 has 3 "pass" scores, while Rater 2 has 4 "pass" scores. So, the product of the number of pass scores is 3 X 4 = 12. Rater 1 has 2 "fail" scores while Rater 2 has one "fail" score. So the product of the number of fail scores is 2 X 1 = 2. Thus, we have 3 X 4 (the product of the number of "1's"), and 2 X 1 (the product of the number of "0's"). Also, we have five total observations, and 5 squared is 25. Thus,

$$P_c = \frac{3(4) + 2(1)}{5 \text{ X } 5} = \frac{12 + 2}{25}$$

$= .56$

and so

$$k = \frac{.80 - .56}{1 - .56} = \frac{.24}{.44} = .55$$

The interpretation of kappa is somewhat different than the other two methods. If the two raters agreed only by chance, kappa would be 0, indicating that there is no agreement between the raters above chance level. A positive kappa index indicates that there is more agreement between the raters than could be expected by chance, while a negative kappa index indicates less than chance agreement between two raters. Kappa, like correlation coefficients, ranges from -1.0 to +1.0.

GLOSSARY

Analytic scoring -- each criterion or component of a task is evaluated individually and assigned a score. Then, the component scores are summed to produce an overall score. It can provide diagnostic information for remediation if that is appropriate.

Blueprint -- a plan for a test which ensures that the test meets its purpose.

Cohen's kappa -- Kappa provides an estimate of agreement between two raters which is corrected for chance agreement.

Concurrent validity -- indicates how well the test predicts current behavior.

Construct validity -- the degree to which the test measures the underlying trait it is supposed to measure. Historically, construct validity was established by looking at the patterns of relationships among item scores or between test scores and other measures. Recently, however, studies have focused on performance differences over time, across groups and settings, and in response to experimental treatments. Both content and criterion related validity are subsumed by construct validity.

Criterion related validity -- based on the degree of empirical relationship between the test scores and criterion scores. This relationship is usually indicated by correlation coefficients or regressions. A distinction is sometimes made between two types of criterion related validity -- predictive validity and concurrent validity.

Graphic rating scales -- words rather than numbers are used as ratings. Graphic rating scales may be quite subjective and may have low reliability if the verbal descriptions are not specific enough. For example, the difference between "neat" and "fairly neat" is subjective and will vary from one person to another.

Halo Effect -- The rater's general impression of a student causes the rater to rate that individual the same on all dimensions.

Holistic scoring -- all criteria are considered simultaneously and only an overall score is assigned. Even with holistic scoring, some criteria can be given more weight than others. Holistic scoring is less time consuming and less expensive but doesn't provide as much information.

Interrater reliability -- How well different raters agree on the score a single performance should receive.

Intraexaminer reliability -- how consistently the rater rates a particular performance across time. It is desirable that the rater's ratings remain constant over time. Factors such as fatigue, comparison effects, etc. may lower the rater's intraexaminer reliability coefficient. This type of reliability should be assessed during rater training.

Logical Error -- A rater error which occurs when the rater believes more or less of a relationship exists between two variables than actually exists.

Numerical rating scales -- rating scales expressed in the form of numbers. Each number should have constant meaning for each task. For example, if "one" is used to indicate unacceptable performance for one task, it should also indicate unacceptable performance for all other tasks.

Performance test -- any assessment procedure which simulates the behavior a student would display on the job with as much realism to the actual job behavior as possible. A performance test will usually, although not always, involve some type of manipulation of equipment, physical response (other than the type required by paper and pencil tests), or oral response.

Personal Bias -- The raters tendency to rate all students at the same point along the rating scale.

Process Scoring -- Scoring of the steps involved in a performance. If the process is to be scored, it must be assessed at the time of performance while a product may be analyzed at a later time. The process may or may not yield a product. Process is often more difficult to assess due to its subjective and/or complex nature. If the essential aspect of the skill is the process, analyze it. Process is usually assessed with a rating scale or checklist.

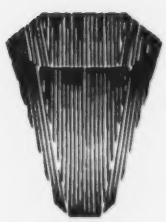
Predictive validity -- indicates how well the test predicts future behavior.

Product Scoring -- The product is a tangible outcome of performance (such as a cake), while process refers to the methods used to arrive at a particular point. If the product is more important than the process, the product should be analyzed. While the product may be assessed with a rating scale, checklist, a product scale (a scale which presents representative samples of products arranged in order of quality and the student's score is assigned on the basis of the product criterion it most closely resembles), or quantitative analysis (e.g., electronic testing).

Reliability -- the degree to which a test measures a characteristic accurately and consistently.

Test-retest reliability -- The same instrument is administered after a period of time has elapsed and the student's scores on the two testings correlated. The period of time elapsed must be long enough to allow for the effects of memory to diminish.

Validity -- Messick (1989) defines validity as
...an integrated evaluative judgment of the degree to which emipirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. p. 13



State of Florida
Department of Education
Betty Castor, Commissioner
Tallahassee, Florida
Affirmative action/equal opportunity employer

Division of Vocational, Adult, and Community Education
Instructional Support Services

AD 434 BK 91

PERFORMANCE TESTING WORKBOOK:

For
Vocational
Education
Administrators
and Teachers

April, 1991

Sponsored and published by:

Florida Deparment of Education
Betty Castor, Commissioner of Education

For further information contact:

Instructional Support Services Section
Bureau of Program and Staff Development
Division of Vocational, Adult, and Community Education
Department of Education
Florida Education Center
Tallahassee, FL 32399-0400
(904) 487-3750/Suncom 277-3750

DVACE/04/91/

TABLE OF CONTENTS

Glossary ---------- iii

I. Introduction ---------- 1

II. Select The Performance Test Development Committee ------ 2

III. Determine What Should Be Tested ---------- 3

IV. Develop A Test Bluprint _____ 5

V. Develop The Performance Test Items _____ 7

VI. Train The Raters ---------- 8

VII. Field Test The Performance Test ---------- 9

VIII. Determine Test Reliability ---------- 10

IX. Determine Test Validity ---------- 11

X. Set Cut Scores ---------- 12

APPENDICES

Appendix A - Task Analysis ---------- 13

Appendix B - Test Blueprint Form ---------- 15

Appendix C - Performance Test Scoring Systems
Performance Checklist ---------- 17
Performance Rating Scale ---------- 18

Appendix D - Sample Performance Test Item ---------- 19

GLOSSARY

Analytic scoring -- each criterion or component of a task is evaluated individually and assigned a score. Then, the component scores are summed to produce an overall score. It can provide diagnostic information for remediation if that is appropriate.

Blueprint -- a plan for a test which ensures that the test meets its purpose.

Cohen's kappa -- Kappa provides an estimate of agreement between two raters which is corrected for chance agreement.

Concurrent validity -- indicates how well the test predicts current behavior.

Construct validity-- the degree to which the test measures the underlying trait it is supposed to measure. Historically, construct validity was established by looking at the patterns of relationships among item scores or between test scores and other measures. Recently, however, studies have focused on performance differences over time, across groups and settings, and in response to experimental treatments. Both content and criterion related validity are subsumed by construct validity.

Criterion-related validity -- based on the degree of empirical relationship between the test scores and criterion scores. This relationship is usually indicated by correlation coefficients or regressions. A distinction is sometimes made between two types of criterion related validity -- predictive validity and concurrent validity.

Graphic rating scales -- words rather than numbers are used as ratings. Graphic rating scales may be quite subjective and may have low reliability if the verbal descriptions are not specific enough. For example, the difference between "neat" and "fairly neat" is subjective and will vary from one person to another.

Halo Effect -- The rater's general impression of a student causes the rater to rate that individual the same on all dimensions.

Holistic scoring -- all criteria are considered simultaneously and only an overall score is assigned. Even with holistic scoring, some criteria can be given more weight than others. Holistic scoring is less time consuming and less expensive but doesn't provide as much information.

Interrater reliability -- How well different raters agree on the score a single performance should receive.

Intraexaminer reliability -- how consistently the rater rates a particular performance across time. It is desirable that the rater's ratings remain constant over time. Factors such as fatigue, comparison effects, etc. may lower the rater's intraexaminer reliability coefficient. This type of reliability should be assessed during rater training.

Logical Error -- A rater error which occurs when the rater believes more or less of a relationship exists between two variables than actually exists.

Numerical rating scales -- rating scales expressed in the form of numbers. Each number should have constant meaning for each task. For example, if "one" is used to indicate unacceptable performance for one task, it should also indicate unacceptable performance for all other tasks.

Performance test -- any assessment procedure which simulates the behavior a student would display on the job with as much realism to the actual job behavior as possible. A performance test will usually, although not always, involve some type of manipulation of equipment, physical response (other than the type required by paper and pencil tests), or oral response.

Personal Bias -- The raters tendency to rate all students at the same point along the rating scale.

Process Scoring -- Scoring of the steps involved in a performance. If the process is to be scored, it must be assessed at the time of performance while a product may be analyzed at a later time. The process may or may not yield a product. Process is often more difficult to assess due to its subjective and/or complex nature. If the essential aspect of the skill is the process, analyze it. Process is usually assessed with a rating scale or checklist.

Predictive validity -- indicates how well the test predicts future behavior.

Product Scoring -- The product is a tangible outcome of performance (such as a cake), while process refers to the methods used to arrive at a particular point. If the product is more important than the process, the product should be analyzed. While the product may be assessed with a rating scale, checklist, a product scale (a scale which presents representative samples of products arranged in order of quality and the student's score is assigned on the basis of the product criterion it most closely resembles), or quantitative analysis (e.g., electronic testing).

Reliability -- the degree to which a test measures a characteristic accurately and consistently.

Test-retest reliability -- The same instrument is administered after a period of time has elapsed and the student's scores on the two testings correlated. The period of time elapsed must be long enough to allow for the effects of memory to diminish.

Validity -- Messick (1989) defines validity as an integrated evaluative judgement of the degree to which emipirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment.

I. INTRODUCTION

Performance testing is defined as many ways as there are authors studying it, yet a common theme purveys all of the definitions. A performance test is best defined as **any assessment procedure which simulates the behavior a student would display on the job with as much realism to the actual job behavior as possible.** A performance test usually involves some type of manipulation of equipment, physical activity, or oral response.

There are a number of steps to follow when developing a performance test. Decisions made throughout the development of the test should be clearly and comprehensively documented. The following steps can be used to develop a performance test:

1. Select a performance test development committee.
2. Determine what should be tested.
3. Develop a test blueprint.
4. Develop the performance test items.
5. Train the raters.
6. Conduct field test.
7. Determine reliability.
8. Determine validity.
9. Set cut scores.

Each of the above steps is discussed in this workbook. Additional in-depth discussion of these topics is provided in the reference manual which is referenced throughout this document.

II. SELECT THE PERFORMANCE TEST DEVELOPEMENT COMMITTEE

For the purpose of developing a performance test for the Gold Seal Vocational Endorsement, form a committee of program advisory committee members, instructors, industry representatives and testing specialists to carry out the steps described herein. State Board of Education Rule 6A-1.092(3d) specifically requires the involvement of the advisory committee in any procedure locally adopted to verify the performance competency of a student for the Gold Seal Vocational Endorsement. Instructors should serve as committee members to provide additional subject-matter expertise and knowledge of resources and constraints of the institution to administer the test.

III. DETERMINE WHAT SHOULD BE TESTED

One criterion of eligibility for the Gold Seal Endorsement is the verification of vocational competency via a performance test. The performance test requires the selection of a representative sample of student performance standards (skills) to be demonstrated by the student. These SPS's must be selected from the current curriculum framework for the vocational program. Several SPS's may be tested by a single performance item. When selecting SPS's for the performance test, the following steps must be taken:

1. Identify the most critical skills from the beginning, middle and end of the program content. Involve your program adivisory committee and instructors.

2. Consider the availability of all resources needed to administer the test.

3. Consider the time available for testing and the time needed to perform the critical skills identified.

4. Select a representative sample of the critical tasks for the performance test.

5. Break down complex skills into their measurable task components. This may be done via an **instructional analysis** of each skill selected for the test (see Appendix A). This information is used to develop the rating scales and or checklists for scoring performance.

(Performance Testing Manual pgs. 2 to 5)

IV. DEVELOP A TEST BLUEPRINT

A test blueprint is the plan for the performance test. The test blueprint helps ensure that the test meets its purpose. The test blueprint contains information about test length, scoring methodologies, weighting of different skills, when to test, the conditions under which testing is to occur, who can serve as a rater, and test security. An example of a test blueprint is provided in Appendix B. The following steps describe how to develop an test blueprint.

1. **Determine test length**
This is determined by the time it takes to perform the Skills selected for the test. The more standards that can be assessed in the test, the more representative it will be. However, it should only be as long as necessary to measure the standards selected. Thus, there may be differences in test length based on the vocational program content.

2. **Determine a scoring system and criteria**
This involves the determination of a rating scale or checklist method for assessing the performance of the skills as observed by the raters. Rating scales require a rater to place the students performance on a continuum. A checklist will require the rater to simply check a skill or subskill when exhibited by the student. An example of both methods is provided in Appendix C.

3. **Assign weights to the Skills**
This reflects the importance and frequency with which tasks are performed on the job and thus translates to the percentage on the overall score of the test. The more important the skill, the higher it may be weighted. Weights are <u>optional</u> and all skills may be assigned equal weight if desired.

4. **Determine when the test will be administered**
For the purposes of the Gold Seal Endorsement, you will be administering an end-of-program test. It is recommended that you test the student after he/she has completed at least 80% of the SPS's for the vocational program, or after program competion (if prior to end of school year).

5. **Identify the conditions for the test**
In order for the test to be standardized among all students, conditions under which the test is to be administered must be clearly identified. It is very important that the conditions applied on the test are consistent with the conditions under which the student training was provided. Conditions for performance include:

a. Directions given to the student
b. Tools and equipment needed to perform the skills
c. Related manuals amd text materials needed to perform the skills
d. Environmental setting for administering the test

6. **Select the performance test rater(s)**
The Gold Seal verification requires the involvement of the program advisory committee in the testing process. Therefore, local industry representatives with specific expertise in the skills being tested are the best source of raters. While teachers may serve as raters, it is recommended that the students' teacher not serve as a rater.

(Performance Testing Manual pgs. 6 to 12)

V. DEVELOP THE PERFORMANCE TEST ITEMS

Once the test blueprint is completed, the performance test may be developed. An example of an actual performance test item is provided in Appendix D. The following steps describe how to develop the test:

1. **Design the test scenario(s)**
Using the information contained in the test blueprint, a real world scenario should be created which requires the student to perform the skills selected for the test. An example of a test scenario is provided in the sample performance test item provided in Appendix D.

2. **Write the instructions**
The instructions to the test administrators, raters, and the students must be written clearly and informatively for each skill that is to be tested. The instructions should state exactly what is expected and the criteria for performance.

3. **List all equipment and materials**
All equipment and materials required for the performance of each part of the test must be listed and included on the test.

4. **Develop rating scales and/or checklists used for scoring**
The system of scoring specified in the test blueprint should be developed. If a rating scale is used, select the number of options on the scale. Regardless of whether a checklist or rating scale is used, all skills to be performed must be clearly identified. The instructional analysis conducted when the skills to be tested were selected will help identify the items for the checklist. An example of both scoring systems is provided in Appendix C.

(Performance Testing Manual pgs. 13 to 18)

VI. TRAIN THE RATERS

Raters must be trained to score the student's performance on the test. The training will include how to use the rating scales and checklists as well as the scoring methodologies used for the performance test. Furthermore, raters must learn how to avoid potential rater biases. The raters must also know the specific skills that the student will be demonstrating on the performance test.

With good rater training, anchored rating scales and checklists, and an analysis of inter-rater reliability, the possibility of inconsistent scoring among raters is greatly reduced.

(Performance Testing Manaul pgs. 18 to 19)

VII. FIELD TEST THE PERFORMANCE TEST

Conducting a field test of the test is highly desirable. This assures that the test has been constructed properly, all instructions to students, raters, and test administrators are clear, the scoring criteria appropriate, the time limits set are reasonable, and estimates of reliability and validity may be obtained. During the field test, administer the items in conditions which will be as close as possible to the conditions under which the items will be administered for the actual test.

(Performance Testing Manual p. 19)

VIII. DETERMINE TEST RELIABILITY

Reliability is the degree to which a test measures a performance standard accurately and consistently. There are two types of reliability most germane to performance testing.

1. **Test-retest reliability**
This is determined by administering the same test at different points in time and correlating the scores. Be sure to allow enough elapsed time to minimize the effects of memory. An example of how to compute a test-retest correlation coefficient is provided in the Appendix in the reference manual.

2. **Inter-rater reliability**
Inter-rater reliability is determined by:
   a. correlating the scores assigned by two or more raters.
   b. dividing the number of agreements by the total number of ratings.
   c. Cohen's kappa.

Reliability can be maximized by:

1. training raters.
2. giving detailed specifications of the task the rater is to perform.
3. providing specific written instructions for raters.
4. reporting information (rating the student) as soon as possible after the observation.
5. using mechanical devises if possible.
6. getting information from several raters who observe at the same time.

Reliability may be determined by:

1. Determine test-retest reliability by administering the same test on two separate occasions and correlating the scores.
2. Determine inter-rater reliability by administering the test and correlating scores assigned by two raters, using Cohen's kappa, or dividing the number of agreements by the total number of ratings.
3. Determine intra-examiner reliability by asking raters to rate the same performance at two points in time and correlating the scores assigned at the different points in time.

(Performance Testing Manual pgs. 19 to 21)

IX. DETERMINE TEST VALIDITY

Validity is the most important feature of a test. Validity can be defined as **the degree to which evidence supports the inferences made or actions taken based on test scores.**

Three categories of validity are most germane to performance tests.

1. **Content validity** may be established by the selection of appropriate items for inclusion in the test.

2. **Concurrent Criterion-related validity** may be established by administering the test to current students and correlating their scores with external criteria such as supervisory ratings.

3. **Construct validity** may be established by determining whether the test discriminates between students who know the content and those who do not.

(Performance Testing Manaul pgs. 21 to 24)

X. SET CUT SCORES

A cut score is the score a student must earn to pass the performance test. Cut scores must be set in a defensible and fair manner, and by persons qualified to make cut score decisions. The ones most qualified to judge cut scores are persons knowledgable in the content area. For example, auto mechanics teachers and auto mechanics themselves should be used to make judgments about auto mechanic test items. The teachers and advisory committee members who determined the test content may be used.

The following steps describe how to determine a cut score.

1. Determine the appropriate cut score methodology.
2. Determine the persons who may serve as judges.
3. Train judges in the cut score methodology selected.
4. Judges set cut score.

(Performance Testing Manaual pgs. 24 to 30)

APPENDIX A

TASK ANALYSIS

Instructional Task Analysis is a procedure that requires both logic and creativity. It takes your knowledge as subject-matter experts to extract the steps of performing a particular job task. It takes creativity to accurately describe those steps.

After you have completed the analysis of a task the simple who, what, when, how check can help assure you have covered most of the information needed.

1. Is **who** does the task relevant and clearly stated or implied?
2. Is **what** is needed to perform the task successfully identified and accurately described?
3. Is **when** the task is applied and completed relevant and clear stated or implied?
4. Is **how** the task is performed described accurately and in appropriate terms?

Procedures for Analyzing Tasks

1. Identify the **Performance skill** from the representative skills selected for the performance test.

2. Determine the **Standard of Performance**, i.e., what level of task performance denotes that it has been satisfactorily completed.

3. Identify the **Conditions for Performance** of the task, e.g., tools, equipment, supplies, environment.

4. Identify step-by-step what a person must know and do in order to perform the task.

5. Develop the **performance rating scale** or **performance checklist** for the task based the analysis of the performance steps of the task.

APPENDIX B

TEST BLUEPRINT FORM

1. Determine Test Length
   a. Number of Tasks _____

   b. Testing Time (hours) _____

2. Scoring

   a. Must student pass <u>every</u> skill? (Yes/No) _____

   b. Are there certain tasks the student <u>must</u> pass? (Yes/No) _____

      Skills:

      _____
      _____
      _____
      _____
      _____

   c. Analytic or holistic scoring? _____

3. Differential Weighting of skills? (Yes/No) _____

	Skill	Points
1.	_____	____
2.	_____	____
3.	_____	____
4.	_____	____
5.	_____	
6.	_____	____
7.	_____	
8.	_____	____
9.	_____	
10.	_____	____

4. When to test? _____

TEST BLUEPRINT FORM - continued

5. Conditions under which testing occurs (test requirements).

_____
_____
_____
_____

6. Rater Selection (Teachers/Industry Representatives)

Rater 1: _____
Rater 2: _____
Rater 3: _____

7. Test security required? (Yes/No) _____
If yes, specifiy security procedures:

_____
_____
_____
_____
_____
_____

APPENDIX C

PERFORMANCE SCORING SYSTEMS

PERFORMANCE CHECKLIST

Directions: Observe the student as he or she performs the task. As the student performs each task, place a check (√) or X in the correct column (Yes or No).

		Yes	No
1.	Measure and correct the ride height to within 1/2 inch of manufacturer's specification.	_____	_____
2.	Read and correct the caster to within 1 degree of manufacturer's specification.	_____	_____
3.	Read and correct the camber to within 1/2 degree of manufacturer's specification.	_____	_____
4.	Read and correct the toe in to within 1/4 degree of manufacturer's specification.	_____	_____
5.	Center the steering wheel to within 2 degrees of manufacturer's specification.	_____	_____

COMMENTS:

Scoring Systems--continued

PERFORMANCE RATING SCALE

Directions: Observe the student as he or she performs each skill. While the student is working, circle the number on the rating scale which assesses the performance of the skill.

PERFORMANCE SKILL - Check Two-wheel Alignment

TASK 1 - Read and adjust the ride height

Error of more than 1.5" from manufac-turer's spec.	Error of 1.1" to 1.5" from manufac-turer's spec.	Error of .5" to 1" from manufac-turer's spec.	Error of less than .5" from manufac-turer's spec.
1	2	3	4

TASK 2 - Read and adjust the caster.

Error of 4 degrees or more from manu-facturer's spec.	Error of 3 degrees from manu-facturer's spec.	Error of 2 degrees from manu-facturer's spec.	Error of 1 degree or less from manu-facturer's spec.
1	2	3	4

TASK 3 - Read and adjust the camber

Error of 3 degrees or more from manu-facturer's spec.	Error of 2 degrees from manu-facturer's spec.	Error of 1 degrees from manu-facturer's spec.	Error of 1/2 degree or less from manu-facturer's spec.
1	2	3	4

Scoring Systems - continued

TASK 4 - Read and adjust the toe-in

Error of 2 degrees or more from manu-facturer's spec.	Error of 1 degrees from manu-facturer's spec.	Error of 1/2 degree from manu-facturer's spec.	Error of 1/4 degree or less from manu-facturer's spec.
1	2	3	4

TASK 5 - Center the steering wheel

Error of more than 10 degrees from center	Error of 6 to 10 degrees from center	Error of 2 to 5 degrees from center	Error of less than 2 degrees from center
1	2	3	4

RATER COMMENTS:

APPENDIX D

SAMPLE PERFORMANCE TEST ITEM

Program: Automotive Mechanics

Performance Objective: Check Two-Wheel Alignment

Scenario: Inspect vehicle on rack and make adjustments needed to wheel system necessary to restore to specifications.

Test Requirements:

Equipment	Supplies	Materials
-safety glasses	-paper towels	-manufacturer's specifi-
-basic hand tools	-fender covers	cations
-alignment rack	-alignment shims	-shop manuals
-alignment equipment		
-specialized alignment tools		

Directions to Test Administrator:

Set up all necessary equipment and supplies in the area where testing will take place (see listed necessary equipment and supplies).

All vehicles must be on the rack and raised to the appropriate height for the student. In addition, all the vehicles' wheel systems should be equally out of alignment.

Provide the raters with:

a. the performance objective

b. the list of necessary equipment and supplies

c. directions to the Raters

d. directions to the Students

e. scoring checklist

Provide extra copies of the Directions to the Students and Scoring Checklist so that all students to be tested may be provided with a copy.

Directions to the Rater:

The student is to inspect and make adjustments needed to the wheel system to restore the vehicle to factory specifications. You are to use the scoring checklist to rate the student's performance as either Acceptable or Unacceptable.

Give the students the Directions to the Student and the Scoring Checklist. Timing of the activity should be as follows:

1. Give the student five minutes to read the **Directions to the Student** and the **Scoring Checklist**. Ask the student if there are any questions. Answer the student's questions.

2. Give the student up to 60 minutes to complete the task.

Directions to the Student:

Your task is to inspect the vehicle on the rack and make any needed adjustments to the wheel system to restore the vehicle to factory specifications.

You have five minutes to read these directions and the Scoring Checklist. The Scoring Checklist indicates how you will be scored on this item.

You have up to sixty minutes to complete the following tasks:
Measure and correct the ride height.
Read and correct the caster.
Read and correct the camber.
Read and correct the toe in.
Center the steering wheel.

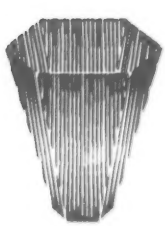
If you have any questions, please ask now.

PERFORMANCE CHECKLIST

Directions: Observe the student as he or she performs the task. As the student performs each task, place a check (✓) or X in the correct column (Yes or No).

	Yes	No
1. Measure and correct the ride height to within 1/2 inch of manufacture's specification.	_____	_____
2. Read and correct the caster to within 1 degree of manufacturer's specification.	_____	_____
3. Read and correct the camber to within 1/2 degree of manufacturer's specification.	_____	_____
4. Read and correct the toe in to within 1/4 degree of manufacturer's specification.	_____	_____
5. Center the steering wheel to within 2 degrees of manufacturer's specification.	_____	_____

COMMENTS:



State of Florida
Department of Education
Betty Castor, Commissioner
Tallahassee, Florida
Affirmative action/equal opportunity employer

Division of Vocational, Adult, and Community Education
Instructional Support Services

AD 433 BK 91